International Journal on Advances in Intelligent Systems

2014 vol. 7 nr. 3&4

The *International Journal on Advances in Intelligent Systems* is Published by IARIA.
ISSN: 1942-2679
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Intelligent Systems
Volume 7, Number 3 & 4, 2014

Ateet Bhalla, Oriental Institute of Science & Technology, Bhopal, India
Freimut Bodendorf, Universität Erlangen-Nürnberg, Germany
Karsten Böhm, FH Kufstein Tirol - University of Applied Sciences, Austria
Pierre Borne, Ecole Centrale de Lille, France
Christos Bouras, University of Patras, Greece
Anne Boyer, LORIA - Nancy Université / KIWI Research team, France
Stainam Brandao, COPPE/Federal University of Rio de Janeiro, Brazil
Stefano Bromuri, University of Applied Sciences Western Switzerland, Switzerland
Vít Bršlica, University of Defence - Brno, Czech Republic
Dumitru Burdescu, University of Craiova, Romania
Diletta Romana Cacciagrano, University of Camerino, Italy
Kenneth P. Camilleri, University of Malta - Msida, Malta
Paolo Campegiani, University of Rome Tor Vergata , Italy
Marcelino Campos Oliveira Silva, Chemtech - A Siemens Business / Federal University of Rio de Janeiro, Brazil
Ozgu Can, Ege University, Turkey
José Manuel Cantera Fonseca, Telefónica Investigación y Desarrollo (R&D), Spain
Juan-Vicente Capella-Hernández, Universitat Politècnica de València, Spain
Miriam A. M. Capretz, The University of Western Ontario, Canada
Massimiliano Caramia, University of Rome "Tor Vergata", Italy
Davide Carboni, CRS4 Research Center - Sardinia, Italy
Luis Carriço, University of Lisbon, Portugal
Rafael Casado Gonzalez, Universidad de Castilla - La Mancha, Spain
Michelangelo Ceci, University of Bari, Italy
Fernando Cerdan, Polytechnic University of Cartagena, Spain
Alexandra Suzana Cernian, University "Politehnica" of Bucharest, Romania
Sukalpa Chanda, Gjøvik University College, Norway
David Chen, University Bordeaux 1, France
Po-Hsun Cheng, National Kaohsiung Normal University, Taiwan
Dickson Chiu, Dickson Computer Systems, Hong Kong
Sunil Choenni, Research & Documentation Centre, Ministry of Security and Justice / Rotterdam University of Applied Sciences, The Netherlands
Ryszard S. Choras, University of Technology & Life Sciences, Poland
Smitashree Choudhury, Knowledge Media Institute, The UK Open University, UK
William Cheng-Chung Chu, Tunghai University, Taiwan
Christophe Claramunt, Naval Academy Research Institute, France
Cesar A. Collazos, Universidad del Cauca, Colombia
Phan Cong-Vinh, NTT University, Vietnam
Christophe Cruz, University of Bourgogne, France
Beata Czarnacka-Chrobot, Warsaw School of Economics, Department of Business Informatics, Poland
Claudia d'Amato, University of Bari, Italy
Sérgio Roberto P. da Silva, Universidade Estadual de Maringá - Paraná, Brazil
Mirela Danubianu, "Stefan cel Mare" University of Suceava, Romania
Dragos Datcu, Netherlands Defense Academy / Delft University of Technology , The Netherlands
Antonio De Nicola, ENEA, Italy
Claudio de Castro Monteiro, Federal Institute of Education, Science and Technology of Tocantins, Brazil
Noel De Palma, Joseph Fourier University, France

Zhi-Hong Deng, Peking University, China
Stojan Denic, Toshiba Research Europe Limited, UK
Vivek S. Deshpande, MIT College of Engineering - Pune, India
Sotirios Ch. Diamantas, Pusan National University, South Korea
Leandro Dias da Silva, Universidade Federal de Alagoas, Brazil
Jerome Dinet, Univeristé Paul Verlaine - Metz, France
Jianguo Ding, University of Luxembourg, Luxembourg
Yulin Ding, Defence Science & Technology Organisation Edinburgh, Australia
Alexiei Dingli, University of Malta, Malta
Mihaela Dinsoreanu, Technical University of Cluj-Napoca, Romania
Ioanna Dionysiou, University of Nicosia, Cyprus
Roland Dodd, CQUniversity, Australia
Nima Dokoohaki, Royal Institute of Technology (KTH)-Kista, Sweden
Suzana Dragicevic, Simon Fraser University- Burnaby, Canada
Mauro Dragone, University College Dublin (UCD), Ireland
Marek J. Druzdzel, University of Pittsburgh, USA
Carlos Duarte, University of Lisbon, Portugal
Raimund K. Ege, Northern Illinois University, USA
Jorge Ejarque, Barcelona Supercomputing Center, Spain
Larbi Esmahi, Athabasca University, Canada
Simon G. Fabri, University of Malta, Malta
Umar Farooq, Amazon.com, USA
Mehdi Farshbaf-Sahih-Sorkhabi, Azad University - Tehran / Fanavaran co., Tehran, Iran
Anna Fensel, Semantic Technology Institute (STI) Innsbruck and FTW Forschungszentrum Telekommunikation Wien, Austria
Stenio Fernandes, Federal University of Pernambuco (CIn/UFPE), Brazil
Oscar Ferrandez Escamez, University of Utah, USA
Agata Filipowska, Poznan University of Economics, Poland
Ziny Flikop, Scientist, USA
Adina Magda Florea, University "Politehnica" of Bucharest, Romania
Francesco Fontanella, University of Cassino and Southern Lazio, Italy
Panagiotis Fotaris, University of Macedonia, Greece
Enrico Francesconi, ITTIG - CNR / Institute of Legal Information Theory and Techniques / Italian National Research Council, Italy
Rita Francese, Università di Salerno - Fisciano, Italy
Bernhard Freudenthaler, Software Competence Center Hagenberg GmbH, Austria
Sören Frey, Daimler TSS GmbH, Germany
Steffen Fries, Siemens AG, Corporate Technology - Munich, Germany
Somchart Fugkeaw, Thai Digital ID Co., Ltd., Thailand
Naoki Fukuta, Shizuoka University, Japan
Mathias Funk, Eindhoven University of Technology, The Netherlands
Adam M. Gadomski**,** Università degli Studi di Roma La Sapienza, Italy
Alex Galis, University College London (UCL), UK
Crescenzio Gallo, Department of Clinical and Experimental Medicine - University of Foggia, Italy
Matjaz Gams, Jozef Stefan Institute-Ljubljana, Slovenia
Raúl García Castro, Universidad Politécnica de Madrid, Spain

Fabio Gasparetti, Roma Tre University - Artificial Intelligence Lab, Italy
Joseph A. Giampapa, Carnegie Mellon University, USA
George Giannakopoulos, NCSR Demokritos, Greece
David Gil, University of Alicante, Spain
Harald Gjermundrod, University of Nicosia, Cyprus
Angelantonio Gnazzo, Telecom Italia - Torino, Italy
Luis Gomes, Universidade Nova Lisboa, Portugal
Nan-Wei Gong, MIT Media Laboratory, USA
Francisco Alejandro Gonzale-Horta, National Institute for Astrophysics, Optics, and Electronics (INAOE), Mexico
Sotirios K. Goudos, Aristotle University of Thessaloniki, Greece
Victor Govindaswamy, Texas A&M University-Texarkana, USA
Gregor Grambow, University of Ulm, Germany
Fabio Grandi, University of Bologna, Italy
Andrina Granić, University of Split, Croatia
Carmine Gravino, Università degli Studi di Salerno, Italy
Michael Grottke, University of Erlangen-Nuremberg, Germany
Vic Grout, Glyndŵr University, UK
Maik Günther, Stadtwerke München GmbH, Germany
Francesco Guerra, University of Modena and Reggio Emilia, Italy
Alessio Gugliotta, Innova SPA, Italy
Richard Gunstone, Bournemouth University, UK
Fikret Gurgen, Bogazici University, Turkey
Maki Habib, The American University in Cairo, Egypt
Till Halbach Røssvoll, Norwegian Computing Center, Norway
Jameleddine Hassine, King Fahd University of Petroleum & Mineral (KFUPM), Saudi Arabia
Ourania Hatzi, Harokopio University of Athens, Greece
Yulan He, Aston University, UK
Kari Heikkinen, Lappeenranta University of Technology, Finland
Cory Henson, Wright State University / Kno.e.sis Center, USA
Arthur Herzog, Technische Universität Darmstadt, Germany
Rattikorn Hewett, Whitacre College of Engineering, Texas Tech University, USA
Celso Massaki Hirata, Instituto Tecnológico de Aeronáutica - São José dos Campos, Brazil
Jochen Hirth, University of Kaiserslautern, Germany
Bernhard Hollunder, Hochschule Furtwangen University, Germany
Thomas Holz, University College Dublin, Ireland
Władysław Homenda, Warsaw University of Technology, Poland
Carolina Howard Felicíssimo, Schlumberger Brazil Research and Geoengineering Center, Brazil
Jingwei Huang , University of Illinois at Urbana-Champaign, USA
Weidong (Tony) Huang, CSIRO ICT Centre, Australia
Xiaodi Huang, Charles Sturt University - Albury, Australia
Eduardo Huedo, Universidad Complutense de Madrid, Spain
Marc-Philippe Huget, University of Savoie, France
Chi Hung, Tsinghua University, China
Chih-Cheng Hung, Southern Polytechnic State University - Marietta, USA
Edward Hung, Hong Kong Polytechnic University, Hong Kong
Muhammad Iftikhar, Universiti Malaysia Sabah (UMS), Malaysia

Prateek Jain, Ohio Center of Excellence in Knowledge-enabled Computing, Kno.e.sis, USA
Wassim Jaziri, Miracl Laboratory, ISIM Sfax, Tunisia
Hoyoung Jeung, SAP Research Brisbane, Australia
Yiming Ji, University of South Carolina Beaufort, USA
Jinlei Jiang, Department of Computer Science and Technology, Tsinghua University, China
Weirong Jiang, Juniper Networks Inc., USA
Hanmin Jung, Korea Institute of Science & Technology Information, Korea
Ilya S. Kabak, "Stankin" Moscow State Technological University, Russia
Eleanna Kafeza, Athens University of Economics and Business, Greece
Hermann Kaindl, Vienna University of Technology, Austria
Ahmed Kamel, Concordia College, Moorhead, Minnesota, USA
Rajkumar Kannan, Bishop Heber College(Autonomous), India
Fazal Wahab Karam, Norwegian University of Science and Technology (NTNU), Norway
Dimitrios A. Karras, Chalkis Institute of Technology, Hellas
Koji Kashihara, The University of Tokushima, Japan
Nittaya Kerdprasop, Suranaree University of Technology, Thailand
Katia Kermanidis, Ionian University, Greece
Serge Kernbach, University of Stuttgart, Germany
Nhien An Le Khac, University College Dublin, Ireland
Reinhard Klemm, Avaya Labs Research, USA
Ah-Lian Kor, Leeds Metropolitan University, UK
Arne Koschel, Applied University of Sciences and Arts, Hannover, Germany
George Kousiouris, NTUA, Greece
Philipp Kremer, German Aerospace Center (DLR), Germany
Dalia Kriksciuniene, Vilnius University, Lithuania
Dariusz Król, AGH University of Science and Technology, ACC Cyfronet AGH, Poland
Markus Kunde, German Aerospace Center, Germany
Dharmender Singh Kushwaha, Motilal Nehru National Institute of Technology, India
Andrew Kusiak, The University of Iowa, USA
Dimosthenis Kyriazis, National Technical University of Athens, Greece
Vitaveska Lanfranchi, Research Fellow, OAK Group, University of Sheffield, UK
Mikel Larrea, University of the Basque Country UPV/EHU, Spain
Angelos Lazaris, University of Southern California, USA
Philippe Le Parc, University of Brest, France
Gyu Myoung Lee, Liverpool John Moores University, UK
Kyu-Chul Lee, Chungnam National University, South Korea
Tracey Kah Mein Lee, Singapore Polytechnic, Republic of Singapore
Daniel Lemire, LICEF Research Center, Canada
Haim Levkowitz, University of Massachusetts Lowell, USA
Kuan-Ching Li, Providence University, Taiwan
Tsai-Yen Li, National Chengchi University, Taiwan
Yangmin Li, University of Macau, Macao SAR
Jian Liang, Nimbus Centre, Cork Institute of Technology, Ireland
Lu Liu, University of Derby, UK
Qing Liu, The Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia
Shih-Hsi "Alex" Liu, California State University - Fresno, USA

Xiaoqing (Frank) Liu, Missouri University of Science and Technology, USA
David Lizcano, Universidad a Distancia de Madrid, Spain
Henrique Lopes Cardoso, LIACC / Faculty of Engineering, University of Porto, Portugal
Sandra Lovrencic, University of Zagreb, Croatia
Jun Luo, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
Prabhat K. Mahanti, University of New Brunswick, Canada
Jacek Mandziuk, Warsaw University of Technology, Poland
Herwig Mannaert, University of Antwerp, Belgium
Yannis Manolopoulos, Aristotle University of Thessaloniki, Greece
Antonio Maria Rinaldi, Università di Napoli Federico II, Italy
Ali Masoudi-Nejad, University of Tehran, Iran
Constandinos Mavromoustakis, University of Nicosia, Cyprus
Zulfiqar Ali Memon, Sukkur Institute of Business Administration, Pakistan
Andreas Merentitis, AGT Group (R&D) GmbH, Germany
Jose Merseguer, Universidad de Zaragoza, Spain
Frederic Migeon, IRIT/Toulouse University, France
Harald Milchrahm, Technical University Graz, Institute for Software Technology, Austria
Les Miller, Iowa State University, USA
Marius Minea, University POLITEHNICA of Bucharest, Romania
Yasser F. O. Mohammad, Assiut University, Egypt
Shahab Mokarizadeh, Royal Institute of Technology (KTH) - Stockholm, Sweden
Martin Molhanec, Czech Technical University in Prague, Czech Republic
Charalampos Moschopoulos, KU Leuven, Belgium
Mary Luz Mouronte López, Ericsson S.A., Spain
Henning Müller, University of Applied Sciences Western Switzerland - Sierre (HES SO), Switzerland
Susana Munoz Hernández, Universidad Politécnica de Madrid, Spain
Adrian Muscat, University of Malta, Malta
Peter Mutschke, GESIS - Leibniz Institute for the Social Sciences - Bonn, Germany
Bela Mutschler, Hochschule Ravensburg-Weingarten, Germany
Deok Hee Nam, Wilberforce University, USA
Fazel Naghdy, University of Wollongong, Australia
Joan Navarro, Research Group in Distributed Systems (La Salle - Ramon Llull University), Spain
Rui Neves Madeira, Instituto Politécnico de Setúbal / Universidade Nova de Lisboa, Portugal
Toàn Nguyên, INRIA Grenobel Rhone-Alpes/ Montbonnot, France
Andrzej Niesler, Institute of Business Informatics, Wroclaw University of Economics, Poland
Kouzou Ohara, Aoyama Gakuin University, Japan
Jonice Oliveira, Universidade Federal do Rio de Janeiro, Brazil
Ian Oliver, Nokia Location & Commerce, Finland / University of Brighton, UK
Michael Adeyeye Oluwasegun, University of Cape Town, South Africa
Sigeru Omatu, Osaka Institute of Technology, Japan
Sascha Opletal, University of Stuttgart, Germany
Fakri Othman, Cardiff Metropolitan University, UK
Enn Õunapuu, Tallinn University of Technology, Estonia
Jeffrey Junfeng Pan, Facebook Inc., USA
Hervé Panetto, University of Lorraine, France
Malgorzata Pankowska, University of Economics, Poland

Harris Papadopoulos, Frederick University, Cyprus
Laura Papaleo, ICT Department - Province of Genoa & University of Genoa, Italy
Agis Papantoniou, National Technical University of Athens, Greece
Thanasis G. Papaioannou, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
Andreas Papasalouros, University of the Aegean, Greece
Eric Paquet, National Research Council / University of Ottawa, Canada
Kunal Patel, Ingenuity Systems, USA
Carlos Pedrinaci, Knowledge Media Institute, The Open University, UK
Yoseba Penya, University of Deusto - DeustoTech (Basque Country), Spain
Cathryn Peoples, University of Ulster, UK
Asier Perallos, University of Deusto, Spain
Christian Percebois, Université Paul Sabatier - IRIT, France
Andrea Perego, European Commission, Joint Research Centre, Italy
Mark Perry, University of Western Ontario/Faculty of Law/ Faculty of Science - London, Canada
Willy Picard, Poznań University of Economics, Poland
Meikel Poess, Oracle, USA
Agostino Poggi, Università degli Studi di Parma, Italy
R. Ponnusamy, Madha Engineering College-Anna University, India
Dorin Popescu, University of Craiova, Romania
Stefan Poslad, Queen Mary University of London, UK
Wendy Powley, Queen's University, Canada
Radu-Emil Precup, "Politehnica" University of Timisoara, Romania
Jerzy Prekurat, Canadian Bank Note Co. Ltd., Canada
Didier Puzenat, Université des Antilles et de la Guyane, France
Sita Ramakrishnan, Monash University, Australia
Elmano Ramalho Cavalcanti, Federal University of Campina Grande, Brazil
Juwel Rana, Luleå University of Technology, Sweden
Martin Randles, School of Computing and Mathematical Sciences, Liverpool John Moores University, UK
Christoph Rasche, University of Paderborn, Germany
Ann Reddipogu, ManyWorlds UK Ltd, UK
Ramana Reddy, West Virginia University, USA
René Reiners, Fraunhofer FIT - Sankt Augustin, Germany
Paolo Remagnino, Kingston University - Surrey, UK
Sebastian Rieger, University of Applied Sciences Fulda, Germany
Andreas Riener, Johannes Kepler University Linz, Austria
Ivan Rodero, NSF Center for Autonomic Computing, Rutgers University - Piscataway, USA
Joel Rodrigues, Instituto de Telecomunicações / University of Beira Interior, Portugal
Alejandro Rodríguez González, University Carlos III of Madrid, Spain
Paolo Romano, INESC-ID Lisbon, Portugal
Agostinho Rosa, Instituto de Sistemas e Robótica, Portugal
José Rouillard, University of Lille, France
Paweł Różycki, University of Information Technology and Management (UITM) in Rzeszów, Poland
Igor Ruiz-Agundez, DeustoTech, University of Deusto, Spain
Michele Ruta, Politecnico di Bari, Italy
Melike Sah, Trinity College Dublin, Ireland
Francesc Saigi Rubió, Universitat Oberta de Catalunya, Spain

Abdel-Badeeh M. Salem, Ain Shams University, Egypt
Yacine Sam, Université François-Rabelais Tours, France
Ismael Sanz, Universitat Jaume I, Spain
Ricardo Sanz, Universidad Politecnica de Madrid, Spain
Marcello Sarini, Università degli Studi Milano-Bicocca - Milano, Italy
Munehiko Sasajima, I.S.I.R., Osaka University, Japan
Minoru Sasaki, Ibaraki University, Japan
Hiroyuki Sato, University of Tokyo, Japan
Jürgen Sauer, Universität Oldenburg, Germany
Patrick Sayd, CEA List, France
Dominique Scapin, INRIA - Le Chesnay, France
Kenneth Scerri, University of Malta, Malta
Adriana Schiopoiu Burlea, University of Craiova, Romania
Rainer Schmidt, Austrian Institute of Technology, Austria
Bruno Schulze, National Laboratory for Scientific Computing - LNCC, Brazil
Wieland Schwinger, Johannes Kepler University Linz, Austria
Hans-Werner Sehring, T-Systems Multimedia Solutions GmbH, Germany
Paulo Jorge Sequeira Gonçalves, Polytechnic Institute of Castelo Branco, Portugal
Sandra Sendra Compte, Polytechnic University of Valencia, Spain
Kewei Sha, Oklahoma City University, USA
Roman Y. Shtykh, Rakuten, Inc., Japan
Kwang Mong Sim, Gwangju Insitute of Science & Technology, South Korea
Robin JS Sloan, University of Abertay Dundee, UK
Vasco N. G. J. Soares, Instituto de Telecomunicações / University of Beira Interior / Polytechnic Institute of Castelo Branco, Portugal
Don Sofge, Naval Research Laboratory, USA
Christoph Sondermann-Woelke, Universitaet Paderborn, Germany
George Spanoudakis, City University London, UK
Vladimir Stantchev, SRH University Berlin, Germany
Claudius Stern, University of Paderborn, Germany
Mari Carmen Suárez-Figueroa, Universidad Politécnica de Madrid (UPM), Spain
Kåre Synnes, Luleå University of Technology, Sweden
Ryszard Tadeusiewicz, AGH University of Science and Technology, Poland
Yehia Taher, ERISS - Tilburg University, The Netherlands
Yutaka Takahashi, Senshu University, Japan
Dan Tamir, Texas State University, USA
Jinhui Tang, Nanjing University of Science and Technology, P.R. China
Yi Tang, Chinese Academy of Sciences, China
John Terzakis, Intel, USA
Sotirios Terzis, University of Strathclyde, UK
Vagan Terziyan, University of Jyvaskyla, Finland
Ioan Toma, STI Innsbruck/University Innsbruck, Austria
Lucio Tommaso De Paolis, Department of Innovation Engineering - University of Salento, Italy
Davide Tosi, Università degli Studi dell'Insubria, Italy
Raquel Trillo Lado, University of Zaragoza, Spain
Tuan Anh Trinh, Budapest University of Technology and Economics, Hungary

Simon Tsang, Applied Communication Sciences, USA
Theodore Tsiligiridis, Agricultural University of Athens, Greece
Antonios Tsourdos, Cranfield University, UK
José Valente de Oliveira, University of Algarve, Portugal
Eugen Volk, University of Stuttgart, Germany
Mihaela Vranić, University of Zagreb, Croatia
Chieh-Yih Wan, Intel Labs, Intel Corporation, USA
Jue Wang, Washington University in St. Louis, USA
Shenghui Wang, OCLC Leiden, The Netherlands
Zhonglei Wang, Karlsruhe Institute of Technology (KIT), Germany
Laurent Wendling, University Descartes (Paris 5), France
Maarten Weyn, University of Antwerp, Belgium
Nancy Wiegand, University of Wisconsin-Madison, USA
Alexander Wijesinha, Towson University, USA
Eric B. Wolf, US Geological Survey, Center for Excellence in GIScience, USA
Ouri Wolfson, University of Illinois at Chicago, USA
Yingcai Xiao, The University of Akron, USA
Reuven Yagel, The Jerusalem College of Engineering, Israel
Fan Yang, Nuance Communications, Inc., USA
Maribel Yasmina Santos, University of Minho, Portugal
Zhenzhen Ye, Systems & Technology Group, IBM, US A
Jong P. Yoon, MATH/CIS Dept, Mercy College, USA
Shigang Yue, School of Computer Science, University of Lincoln, UK
Constantin-Bala Zamfirescu, "Lucian Blaga" Univ. of Sibiu, Romania
Claudia Zapata, Pontificia Universidad Católica del Perú, Peru
Marek Zaremba, University of Quebec, Canada
Filip Zavoral, Charles University Prague, Czech Republic
Yuting Zhao, University of Aberdeen, UK
Hai-Tao Zheng, Graduate School at Shenzhen, Tsinghua University, China
Yu Zheng, Microsoft Research Asia, China
Zibin (Ben) Zheng, Shenzhen Research Institute, The Chinese University of Hong Kong, Hong Kong
Bin Zhou, University of Maryland, Baltimore County, USA
Alfred Zimmermann, Reutlingen University - Faculty of Informatics, Germany
Wolf Zimmermann, Martin-Luther-University Halle-Wittenberg, Germany

International Journal on Advances in Intelligent Systems
Volume 7, Numbers 3 & 4, 2014

CONTENTS

Interactive Lectures: Encouraging Student Interaction Using Quick Response Codes

Robert Law
Computer, Communications and Interactive Systems
Glasgow Caledonian University
Glasgow, Scotland
e-mail: robert.law@gcu.ac.uk

Abstract— **This article presents an ongoing project to encourage student interaction during lectures through the use of Quick Response (QR) codes and Google forms to generate rapid response polls and quizzes. Audience Response Systems (ARS) are generally expensive to purchase, require students to purchase a clicker and need to be installed in a dedicated room. This article proposes the use of Google applications (Apps) software combined with the students' own smartphone as a viable free alternative to the current clicker systems. An investigation into the pedagogical issues associated with such a project will be explored and an attempt made to incorporate these into the student experience. The overall process from the creation of the software to the roll out and use of the software in an interactive lecture, the issues encountered and participant feedback will also be described.**

Keywords- QR Codes; Student Interaction; Feedback.

I. INTRODUCTION

The initial paper of this work [1] outlined the combination of readily available, and to certain extent, free technologies that could be combined to enhance the students' and lecturers lecture experience.

Smartphones in the UK, and elsewhere, have seen a surge in popularity over the last few years evidenced by recent figures showing "Over half of the British population (50.3%) now owns a smartphone" [2]. A survey conducted by Deloitte [3] in the UK during the early part of 2014 points to a quite considerable rise in smartphone ownership. Deloitte's survey indicates that smartphone ownership is up 10% between 2013 and 2014 leading to almost 80% of households in the UK owning at least one smartphone. This surge in ownership strengthens the belief that the smartphone has become a ubiquitous technology.

Edinburgh University recently conducted a survey of their student population determining that 67% had ownership of a smartphone "an increase of seventeen percent from those students surveyed seven months earlier" [4]. This uptake in smartphone ownership within the student population opens a new dimension for interaction.

A tangible increase and use of Quick Response (QR) codes by many companies as a form of marketing has ensued on the back of this increase in smartphone popularity. This has allowed many companies to develop new and engaging ways for their customer base to interact with products in situ reinforcing brand presence. This potential has already been harnessed by Education to extend learning materials through the use of QR codes. Learning materials have been enhanced by providing "just in time support materials" [5] such as videos, explanatory text, Uniform Resource Indicators (URI) and staff details.

Using this as a platform to build from the next logical step is to combine the technologies to allow students to interact during lectures through quick multiple choice based questions. The students' responses can be compiled to show the result in a timely manner [6].

The remainder of this article is organized as follows: Section II gives information about pedagogical issues related to interactive lectures, Section III introduces the technology used for implementing interactive lectures; this covers both hardware and software. Section IV discusses the author's experience of implementing interactive lectures, while Section V discusses issues encountered during the interactive lectures. Section VI discusses selected feedback garnered from a student survey. Section VII offers a summary of our experience of interactive lectures, and concludes the article giving proposals for future work.

II. PEDAGOGICAL ISSUES

Take a typical lecture; what does this encompass? Information is imparted upon the student in a relatively one way passive communication format. This traditional didactic approach is a format that has been used for centuries. This research intends to explore the possibility of improving and enhancing the lecture experience through the use of technology, and in particular, Audience Response Systems (ARS). The ability of such systems to encourage active learning through student participation and engagement provides an opportunity for enhancing the passive lecture format by introducing two way interactions with the student audience [9].

Murphy and Sharma further suggest that the research literature available for the topic of interactive lectures and the related pedagogical issues are "almost non-existent, with major issues waiting to be examined… inadequate research on the pedagogical implications of the emerging interactive

forms of learning" [9]. With this in mind there appears to be an opportunity to examine and suggest how ARS technology could be used to not only enhance lecturer-student interactions, but also develop the underlying pedagogical issues inherent with lectures.

The concept of "Deep and lasting learning" as postulated by Boyle and Nicol [10] is enhanced when students have the ability to "actively engage" with what they are learning, allowing them to "construct their own understanding".

The theory of punctuating a lecture is not new as evidenced by Angelo and Cross [11] and the concept of the "minute paper" whereupon the student body is asked to answer the question "What was the most important thing you learned during this class?". The students take the last few minutes of the lecture to answer this question giving the lecturer important written feedback. This method allows the lecturer to gage how well the students have understood the delivered content but the timeliness of the feedback is not effective for the student.

ARS technology provides a means for the lecturer to engage and interact with the students using the responses to offer the student audience immediate feedback. This should lead to further discussion and the opportunity for student reflection. Other research, reported in Murphy and Sharma [9], identifies two pedagogical aspects of interactive lecturing: dialogic form of learning and active learning.

The project intends to examine these issues and the resultant effects that they have on the student audience. The primary concern is that the interactive lecture will stimulate engagement and interaction with the student audience and the lecturer through the use of instant feedback. This feedback will engender in both the student audience and the lecturer the need for reflection on many aspects of the material delivered and possibly the module in general. Through the use of relevant and targeted questions the students can be cajoled into discussions that will help expand their understanding of the topic area. Through these discussions both the students and the lecturer will be able to better understand the level of the students understanding of the subject area.

Gannon-Leary et al. [7] reported a number of other positive aspects to arise from interactive lectures including improved concentration, greater enjoyment and improved attendance. Simpson and Oliver [14] also noted that the anonymity provided by ARS technology played an important part in encouraging students to contribute to answering questions but suggests, as does Gannon-Leary et al. [7], that the design of the questions is very important to the process.

Saravani and Clayton [15] have developed a conceptual framework referred to as A.C.E. This framework is composed of the three A's: Awareness, Action, and Accomplishment; three C's: Context, Content, and Capability; and the three E's: Enabled, Engaged, and Empowered. The three E's aspect of the framework fits the concept driving interactive lectures as the use of mobile technology enables, engages and empowers both the student body and the lecturer.

Devon and Law [21] have mooted a framework of five stages which they believe are required "to maximize the benefit of the technology".

Table I shows the framework noting the "Process", the participant "Who" and the point at which each step takes place. This table highlights the need to prepare the question prior to the lecture. The preparation of a carefully crafted question will maximize the benefit of its use for both the student's understanding and the lecturer's ability to ascertain relevant feedback.

TABLE I. DEVON AND LAW FRAMEWORK

Process	**Who**	**When**
BUILDing the question	Lecturer	Before
ASKing the question	Lecturer/Student	During
CONSIDERing the question	Student	During
DISCUSSing the question	Lecturer/Student	During
EVALUATEing the question	Lecturer/Student	After

The framework proposed by Devon and Law [21] is influenced by Beatty et al. [22] question driven instruction (QDI). QDI suggests that the process of crafting a question that elicits the maximum benefit for the learner, asking this question, allowing the learner time to process the question and answer the question before facilitating a discussion based around the answers given is a "key vehicle for learning".

Sandhu, Afifi, and Amara [23] assert the need to grab the learner's attention and subsequently hold their attention is a key aspect of any interactive lecture. They continue to propose that questioning, discussion and timely feedback are core to effective lectures.

A common theme highlighted by a number of researchers [7] and [14] is the need for well-designed questions and their correct placement within the lecture. If the questions are poorly construct and or badly placed within the lecture this will negate any perceived benefit from the use of the interactive system.

Ramsden [5] suggests, amongst other things, that Quick Response (QR) codes can be used for "just in time information in a face to face lecture"; drawing on this point allows for the expansion of the concept to include feedback for both the student and the lecturer. Mooted by Law and So [17] is the idea that QR codes can facilitate a "trinity of "location independence," "time independence" and "meaningful content"". Of interest in this "trinity" is the idea of "location independence"; being able to deliver and receive feedback to and from the students in the lecture hall.

Both Dufresne et al. [12] and Crouch and Mazur [13] indicate the use of interactive lecture systems, whether as clickers or the system proposed here, can help facilitate opportunities for in lecture student discussion and as such improve the students' retention and understanding of the topic delivered.

III. TECHNOLOGY

ARS systems are available in many forms and price points. A typical classroom package can cost $1000 or more

for software, receiver and 12 clickers [18]. Some systems require the student to purchase a clicker and yearly registration at a cost of $20/$15, respectively [17]. This project developed a "no cost in-house" system that was based on three key components: smartphones, QR codes and a Google spreadsheet. No proprietary software for the phone is required simply a standard browser and bar code reader. The "back-end" is relatively simple to create as the implementation interface is supplied by Google.

A. Hardware

The student participants referred to in Section IV were surveyed to determine the spread of handset manufacturers and phone operating systems. Around 50 students were surveyed.

Figure 1 shows the spread of handset manufacturers and Figure 2 shows the spread of phone operating systems with the surveyed population.

With such a range of manufacturers and operating systems an "app" based solution would be time consuming and prohibitive. With further investigation, it was discovered that third party barcode reading software was available for all the platforms, thus allowing the students to use their own phones for participation in the lectures.

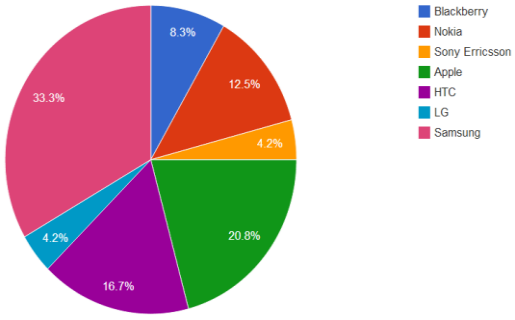


Figure 1. Spread of manufacturers

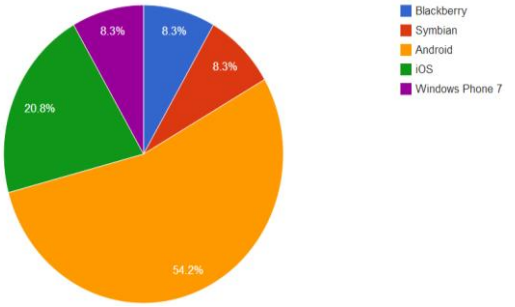


Figure 2. Spread of operating systems

B. Software

The software can be split into three categories: third party barcode reading software, third party browser software and the development of the interactive engine software using Google spreadsheets. It was up to the students themselves to decide on a suitable third party barcode reading software, although some phones did have such software preinstalled.

Three components of Google spreadsheets were used in the creation of the software application: the spreadsheet, the form and Google Script.

The spreadsheet itself is used as a repository for the student responses and also to house a summary sheet which keeps a running total of the number of responses for each possible answer to the question. The work horse of the system is the form and the scripts generated to process the responses at the back end. When a spreadsheet is created using Google Drive, a unique identifier is generated to identify the spreadsheet. When the subsequent form is created for the spreadsheet, another unique identifier is generated.

Although Google forms can handle a number of different inputs, the decision was taken to keep the question to a simple multiple choice style question, thus presenting the student with two or three QR codes per question. To create the QR codes requires the compilation of an http request based on the URI for the Google spreadsheet and the data to be sent to the spreadsheet. Once the http request was constructed and tested an online QR code generator was used to generate the required QR codes. These QR codes were subsequently saved as image files for insertion into the lecture slides.

Google script was used to create code that processed the student responses as they were received to generate a response summary that was visible to the student audience.

C. Code Explanation

This section will provide an examination of the code used to build the application. Before coding the application began an understanding of how Google compiles and uses query strings to transfer data between the application and its servers is required. The application hinges on the query string being built correctly.

Automating the process to build the QR code means that no prior knowledge of how to use or create QR codes is required, thus, widening the access of the tool.

Google provides a number of Application Programming Interfaces (api's), one of which, is the chart api. This api is accessible in a number of different ways, but for this application a direct call will be made to the chart api which is located at Google.com.

A bit of reverse engineering is required to understand the composition of the Uniform Resource Locator (URL) generated by entering data in the Google form. An example URL is shown in Figure 3.

The string contained in the spreadsheet at cell C6 uses the spreadsheet, concatenate function, to combine a number of strings and data from different cells. The first part of the string that forms the URL, "https://docs.google.com/spreadsheet/formResponse?formkey=", has two distinct parts. The "https://docs.google.com/" refers to the location that Google stores it's "docs" suite of apps. The "spreadsheet/" refers to the particular application from the suite that will be used. The next part is important,

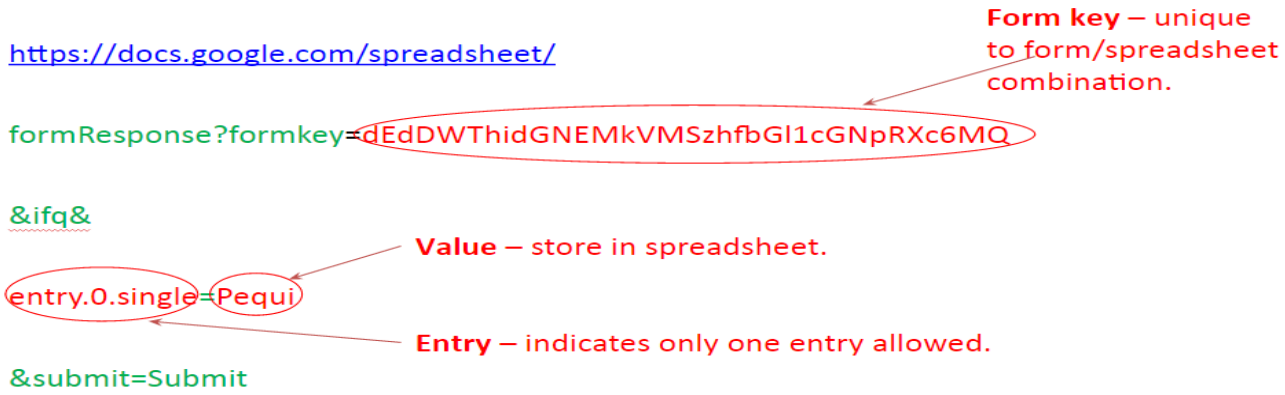


Figure 3. Example URL

this is the query string that tells Google the required processing that its server needs to perform. The "formResponse?formkey=" indicates that the subsequent data following the question mark comprises the response to a form. This is the data that would be generated and sent if the user had to fill in a Google Form. The first parameter after the question mark is "formkey"; this is vitally important as it refers to the unique identifier given to the spreadsheet by Google when it was created. This value is stored in cell B1. Each parameter is always separated by an ampersand. The next parameter "&entry.0.single=" indicates that a single entry of data is required. This means that the data has been limited to a single piece of information. The value of this data is stored in cell B6. The final parameter "&submit=Submit" is used to signify that the data entry is complete and is now ready for processing. The URL generated by this process will subsequently be used to generate a QR code that the user will scan using their phone.

The composition of the formula to produce the QR Code is based on the nesting of spreadsheet commands.

```
=image(CONCAT("http://chart.apis.google.com/chart?cht=qr&chs=350x350&chl=",encodeURL(C6)))
```

Figure 4. Example Spread sheet function to create QR Code.

The CONCAT command is used to join two strings together. The first string is the reference to the Google chart api; it is a parameterized string containing the minimum parameters required by Google to correctly generate the specified chart type. Breaking down the string, "chart.apis.google.com" points to the location online where Google stores it's chart api code. The following part of the string specifies the parameters required chart for generating the desired chart type. The "/chart?" indicates that a chart is required. The parameters that follow make up the response string that will indicate the desired chart type, size and data to be encoded in the qr code, each separated by the use of an ampersand. The "cht=qr" indicates that the chart type will be QR. The "chs=350x350" indicates the size of the QR code to be generated. In this instance the size will be 350 pixels by 350 pixels. The "chl=" indicates the data that will be encoded as part of the QR code. This will be determined by the concatenation of the second string. The second string is contained the spreadsheet at cell C6. The string stored at this location will be used as a web address; it is advisable, therefore, to use the function encodeURL to ensure the correct encoding of the string as this will have a crucial affect when the QR code is subsequently scanned and an attempt to access the web address is made.

The scripts for the prototype are relatively simple in their construction. The example in Figure 5 shows the use of six global variables and a controlling method called runChart().

```
1  // Variables accessable by all functions in the code script
2  var sheet = SpreadsheetApp.getActiveSpreadsheet();
3  var sheetActive = SpreadsheetApp.getActiveSheet();
4  var dataSheet = sheet.getSheets()[0];
5  var summarySheet = sheet.getSheets()[1];
6  var summaryRange = summarySheet.getRange("a1:b5");
7  var chart = sheetActive.getCharts()[0];
8
9  function runChart(){
10   if ( sheetActive.getCharts().length == 0)
11   {
12       createChart_();
13   }
14   else
15   {
16     updateData_();
17     modifyChart_();
18   }
19 }
20
```

Figure 5. Variable Declarations and Controlling Method.

The global variables are used to make the code more concise. These variables rely on Google script commands to access various aspects of the spreadsheet. The first variable on line 2 determines the active spreadsheet; the second variable on line 3 determines the active sheet within the active spreadsheet. The variables on lines 4 and 5 are used to target specific sheets in the active spreadsheet. Line 4 refers to the sheet containing the data returned by scanning

the QR code and line 5 refers to the summary sheet used to keep track of the running totals for each possible response. See Figure 7 for an example of possible data.

Line 6 looks at a specific range of cells on the summary sheet to find out the final totals for each possible response. Line 7 sets the sheet that will be used to draw the chart.

Lines 9 to 19 form the controlling method that is activated each time a response is received. This method checks for the presence of an existing chart, if no chart is present then a new chart will be generated by making a call to the method createChart(). The code for this method is shown in Figure 6.

```
57 // Local function to create and insert a new chart
58 function createChart_() {
59
60   var chart = sheetActive.newChart()  //.newColumnChart
61     .setChartType(Charts.ChartType.COLUMN)
62     .addRange(summaryRange)
63     .setOption('title', 'Votes - Last updated ' + new Date().toString())
64     .setOption('legend', {position: 'right'})   // , title: 'Votes'
65     .setOption('legend', {title: 'Votes'})
66     .setOption('hAxis', {title: 'Answers'})
67     .setOption('vAxis', {title: 'Votes'})
68     .setPosition(3, 3, 0, 0)
69     .build();
70   sheetActive.insertChart(chart);
71 }
72
```

Figure 6. createChart() method.

The code in Figure 6 makes use of Googles newChart() builder command to create a new chart to be displayed on the active sheet. The command takes as parameters the type of chart to be built, the data to be used to build the chart and a number of optional parameters used to configure titles, labels, and axes. The last command in the method, line 70, is used to insert the newly created chart in the active sheet.

If the chart is already in existence then a call is made to the method updateData(). This method is used to recalculate the running totals on the summary sheet.

```
21 // check and update the totals
22 function updateData_() {
23
24   var lastRow = sheet.getLastRow();
25   var dataRange = "a2:b" + lastRow;
26   var chartData = dataSheet.getRange(dataRange);
27
28   var iPequi = 0;
29   var iPupunha = 0;
30   var iPineapple = 0;
31   var iPawpaw = 0;
32
33   for (var iRow = 1; iRow < lastRow; ++iRow)
34   {
35     var dataCell = "b" + iRow;
36
37     var dataValue = chartData.getCell(iRow, 2).getValue();
38
39     switch (dataValue)
40     {
41         case 'Pequi': iPequi++;break;
42         case 'Pupunha': iPupunha++;break;
43         case 'Pineapple': iPineapple++;break;
44         case 'Pawpaw': iPawpaw++;break;
45     }
46
47  }
48
49   // update summary table
50   summaryRange.getCell(2, 2).setValue(iPequi);
51   summaryRange.getCell(3, 2).setValue(iPupunha);
52   summaryRange.getCell(4, 2).setValue(iPineapple);
53   summaryRange.getCell(5, 2).setValue(iPawpaw);
54
55 }
56
```

Figure 7. Method for updating summary sheet.

Figure 7 shows the code used to recalculate the summary sheet values. In summary the process involves traversing each row of the data sheet determining the response and updating the running total accordingly.

The number of rows in the data sheet will vary by the number of responses received and as such will be different each time the method is activated. Google script provides a command, getLastRow(), which enables the method to determine the last row in the sheet (Figure 7, line 24). Line 25 shows how the data range can subsequently be built by combining the last row with the string "a2:b". This can then be used to grab the data range for processing (Figure 7, line 26).

Lines 28 to 31 simply set the summary counters to zero to ensure that the totals are correctly determined. Lines 33 to 47 implement a for loop that checks the contents of each cell in the data range and increments the corresponding summary total. Lines 50 to 53 then update the summary sheet with the newly calculated totals.

After the summary sheet has been recalculated the chart must be redrawn to reflect the new data. The code in Figure 8 makes use of the chart modify command to update the chart.

```
73 // Local function to modify an existing chart.
74 function modifyChart_(){
75   chart = chart.modify()
76               .removeRange(summaryRange)
77               .addRange(summaryRange)
78               .setPosition(5, 5, 0, 0)
79               .setOption('title', 'Last updated ' + new Date().toString()) // Update title.
80               .build(); // Must be called to save changes.
81
82   sheetActive.updateChart(chart);
83
84 }
```

Figure 8. modifyChart() Method.

Essentially this method removes the previous data range from the chart before adding the new, updated, data range to the chart. These updates are then built before the Google command, updateChart(chart), is issued to update the existing chart.

D. Process

The students used their smartphones to scan the QR code of their choice, submitting the data request via their phones browser. This, in turn, populated the spreadsheet with the students' choices, activating the script, allowing the results to be observed in near real time.

The QR Code is an encoded representation of the Google URI and data that will be sent to the spreadsheet when student scans it. As can be seen in Figure 9 below, once scanned, the information encoded in the QR code is decrypted and becomes visible to the phone's user. At this point, the participant has the ability to accept or decline the invite to send the data request. When the participant accepts the request to send the data, the next step is to invoke the phone's browser (this can work in different ways, depending on the phone/operating system), which will send the http request to Google for processing. Once processed, a "thank you" message is displayed in the browser indicating the data request has been received.

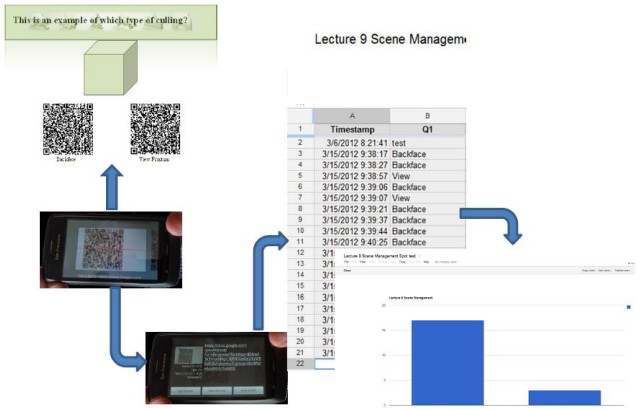


Figure 9. The interactive process

Once the data request has been received, the data is placed in the spreadsheet; the data will be based on the URI encoded in the QR Code. An example is shown below in Figure 10.

It is noticeable from Figure 10 that the spreadsheet date and time stamps the entries as it receives them. The only data that the spreadsheet records is the data encoded in the QR Code and the date and time stamp it generates on receipt of the data. Hence, all entries are anonymous.

	A	B
1	Timestamp	Which one is true?
2	3/22/2012 9:25:08	DAG
3	3/22/2012 9:26:13	DAG
4	3/22/2012 9:26:17	DAG
5	3/22/2012 9:26:20	DAG
6	3/22/2012 9:26:25	DAG
7	3/22/2012 9:26:33	BST
8	3/22/2012 9:27:01	DAG
9	3/22/2012 9:28:00	DAG
10	3/22/2012 9:28:01	DAG
11	3/22/2012 9:29:18	DAG
12	3/22/2012 9:33:01	DAG
13	3/22/2012 9:34:50	Octree
14	3/22/2012 9:35:43	Two
15	3/22/2012 9:35:47	Two
16	3/22/2012 9:35:53	Quadtree
17	3/22/2012 9:37:15	Two
18	3/22/2012 9:42:17	Two

Figure 10. Data sent to the spreadsheet

fx Rast

	A	B
1	Rast	13
2	Geo	6
3	Vertex	0
4	Stencil	12

Figure 11. Summary sheet

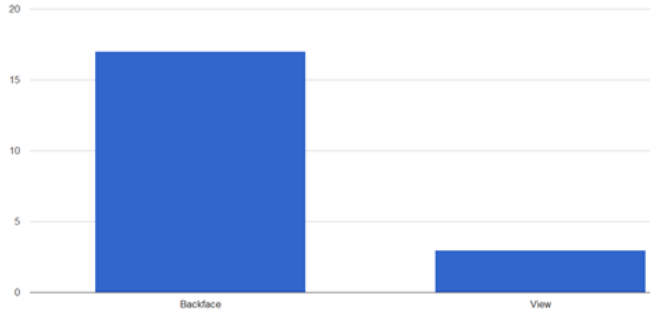


Figure 12. Generated bar chart

As each entry is received, a script is triggered, which counts the entries based on the predefined data set and populates a summary sheet which is used to generate the "near real time" bar charts.

An example of a summary sheet is shown in Figure 11 and an example of the bar chart displayed to the students is shown in Figure 12.

IV. EXPERIENCE

A set of initial tests were developed to examine the viability of the technology and gauge the reaction of the students to the use of this technology within the lecture environment. The aim of the tests was to introduce the interaction concept in a gradual staged manner that would not over burden the student or detract from the lecture.

A. *Test Setup*

The process used to create this interactive lecture was based on designing a set of suitable questions that could be used to strategically punctuate the lecture to gain maximum benefit for the students [17] [22].

The desired effect was to integrate the technology within the lecture while stimulating interaction with the students [17]. As such, the first set of tests was built to increase in a systematic manner the number of questions within the lecture and the number of QR codes within the questions.

The structure of the first set of tests was designed to build from one question with one QR code in the lecture to four questions with three codes per question in the lecture. Suitable points within the lecture were identified such that the questions could be inserted to maximise their impact. An attempt was made to define suitable points through natural break points within the lecture, e.g., end of a topic, end of the lecture, worked example. Using this principal, questions could be deployed with the aim of giving both the student and the lecturer instant feedback on the comprehension of the material delivered.

B. *Participants*

Both sets of participants were studying on the Games Software Development Degree. The first group to undertake the interactive lectures was a second year cohort of around 30 students and the second group to undertake the interactive lectures was a final year cohort of around 20 students.

The second year cohort had three consecutive lectures. The first of the three lectures had one question with two QR codes positioned at the end of the Lecture. The second Lecture had three questions, each with two QR codes positioned at appropriate points within the Lecture and the final Lecture had four questions, each with three QR codes again positioned at appropriate points within the Lecture.

The final year cohort had two lectures which were non-consecutive. The first of the two lectures had two questions each with three QR codes positioned at appropriate points within the Lecture and the second Lecture had two questions each with three QR codes positioned at appropriate points within the Lecture.

At the appropriate point in the lecture, the slide would be displayed. To help minimize issues with scanning, article copies of the slide were also distributed. This allowed for the difference within the quality of phone cameras to focus on the projected QR codes. It was fully explained to the students the nature of the experiment and the procedure which should be followed to correctly participate in the experiment.

C. *Feedback*

Initial feedback from both test groups has been positive and very informative. Feedback ranged from the ease of operation of the process to the size of the QR codes. In general a "buzz" was created within the participant groups generating a positive reaction from the students. This reaction must be tempered by the fact that the students are open to the "Hawthorne effect" [20].

In order to ascertain a clearer picture of the systems use and acceptability with students subsequent testing was planned and undertaken.

It was decided that the most consistent course of action would be to replicate the tests outlined in Section IV B with another Year 2 and Year 4 cohort.

V. ISSUES

Although the overall outcome of the first set of experiments was positive, a number of issues were highlighted that require further polishing prior to the next set of experiments being run.

This section will be divided into two subsections dealing with issues that are prevalent to students and to staff.

A. *Issues related to Students*

An issue flagged up by the participants centered on the size and positioning of the QR codes as this can have an impact on the accuracy of the scanning process.

The student participants indicated that the size of the QR codes on a projected slide proved difficult to scan directly. Not all students were able to scan it directly. This had been anticipated and article based copies of the slides were issued to counter that problem.

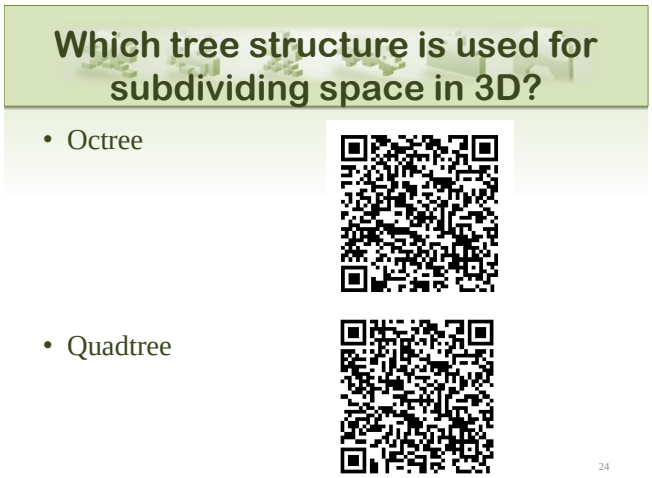


Figure 13. Two QR Code Stack

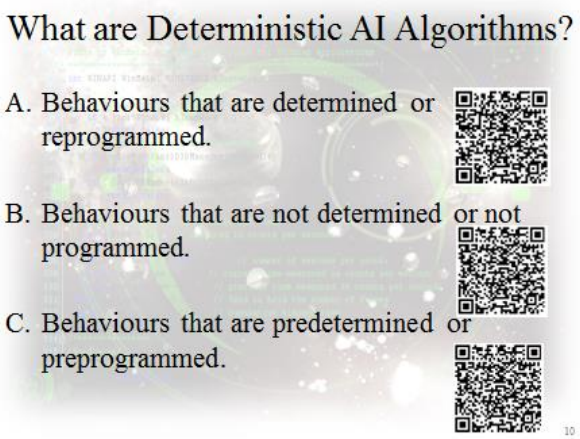


Figure 14. Three QR Code Stack

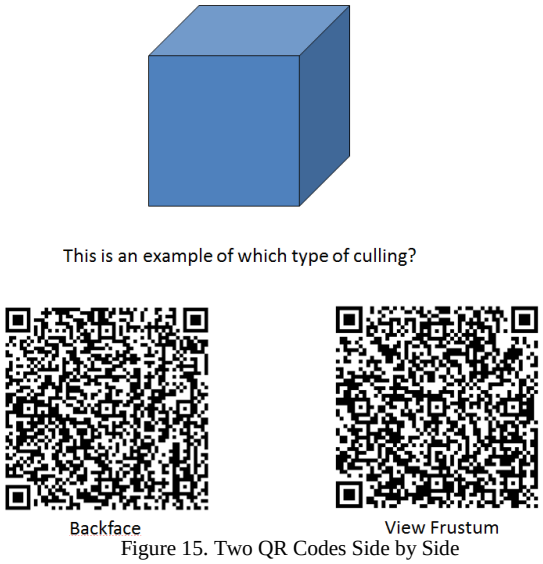


Figure 15. Two QR Codes Side by Side

Figure 16. Three QR Codes Side by Side

Positioning of the QR Codes on the slide raised debate with the participants as some indicated that the barcode scanner software could find it difficult to focus on the required code. Figures 13 to 16 illustrate a number of variations tried out in order to find a balance between size and positioning.

When the question posed had only two available choices there was more leeway to arrange the text and QR codes for both on screen and printed formats. Dealing with two QR codes was relatively straight forward with regard to size and positioning although the student preference was for the stacked vertically display.

Dealing with three QR codes was more problematic as can be seen between the textual difference seen in Figure 14 and Figure 16. Questions that required a more substantial descriptive answer were neigh on impossible to position in a usable manner. The number of QR codes on the slide also influenced the ease of scanning, with three codes per slide proving more challenging for the phone's barcode scanner. This was not insurmountable, but merely added a small time overhead as the participants positioned their camera phone to optimize the scan. Again students opted for the stacked vertically approach.

The overarching problem that affected scanning the QR codes was the quality and abilities of the smartphones camera to focus on the codes.

Camera focusing on the paper based codes generally was satisfactory but a small number of cameras had problems focusing when the QR Codes were small and closely grouped together.

In large lecture theatres none of the students could manage to scan the codes directly from the slides. However, in the small lecture theatres a number of students who had iPhone 5's were able to scan the codes directly. This appeared to be related to the cameras ability to zoom in focusing on a particular area.

B. Issues related to Staff

Time management of both the interactions and the subsequent discussions should be built into the lecture timings allowing for leeway should anything go awry with the technology. Although the technology is improving it is always worth allowing for small technical glitches and having a cutoff point at which no more tinkering will be done to make the phone scan the QR code.

Some of the issues that have become apparent while trialing the system are not insurmountable but can eat into the time available.

The most common issue faced was the number of students who, on the first exposure to the system, did not have a QR code reader installed on their phone. On the very first run on the system a mistake assumption was made that all student's phone's would have QR code readers and as such time was not allocated for the downloading and configuring of the QR code software on the phones. To a certain degree this can be negate by asking all students to have the relevant software installed on their phones prior to the lecture.

Another technical issue that can cause consternation and affect the resultant graph being displayed is the possibility of multiple scans by the occurring from the same device.

This can be due to the barcode scanner software the students were using. Students should be advised to check the settings of the QR Code software that they are using and where possible set the software to require an acknowledgement or conformation prior to sending the scanned web address.

On one or two occasions during the trial of the system the graph being displayed did not update itself as quickly as might have been expected.

Refreshing the chart manually will help reset the graph but it does have an effect on the desired impact of the interactivity of the lecture.

This approach relies on all the students in the lecture having a smartphone and it is conceivable that a small percentage of the student audience may in fact not have a smartphone or even a mobile phone. The perception would be that this student would be disadvantaged by not being able to take part in the interaction. This would be, of course, true. However, it could be mooted that the student is still engaged in the wider discussion that will come from viewing the generated graph. Another possibility is to rely on the goodwill of a fellow student to share access to their phone.

Currently, the whole process required to create and integrate the QR codes into a lecture are quite cumbersome and may prove challenging for a non-computing subject specialist. Since the first trial run outlined in the original paper [1] the system has been significantly restructured to be more user friendly and have a better user interface. The system now has a menu based approach to allow the lecturer to create the question and answers they would like to use.

Once the question and answers have been created and submitted the automated process creates a Google form, Google spreadsheet and the QR codes.

The Google form will be used to collect the student's submissions via the QR code scanned and the spreadsheet with attached script will process the data to produce the near real time graph.

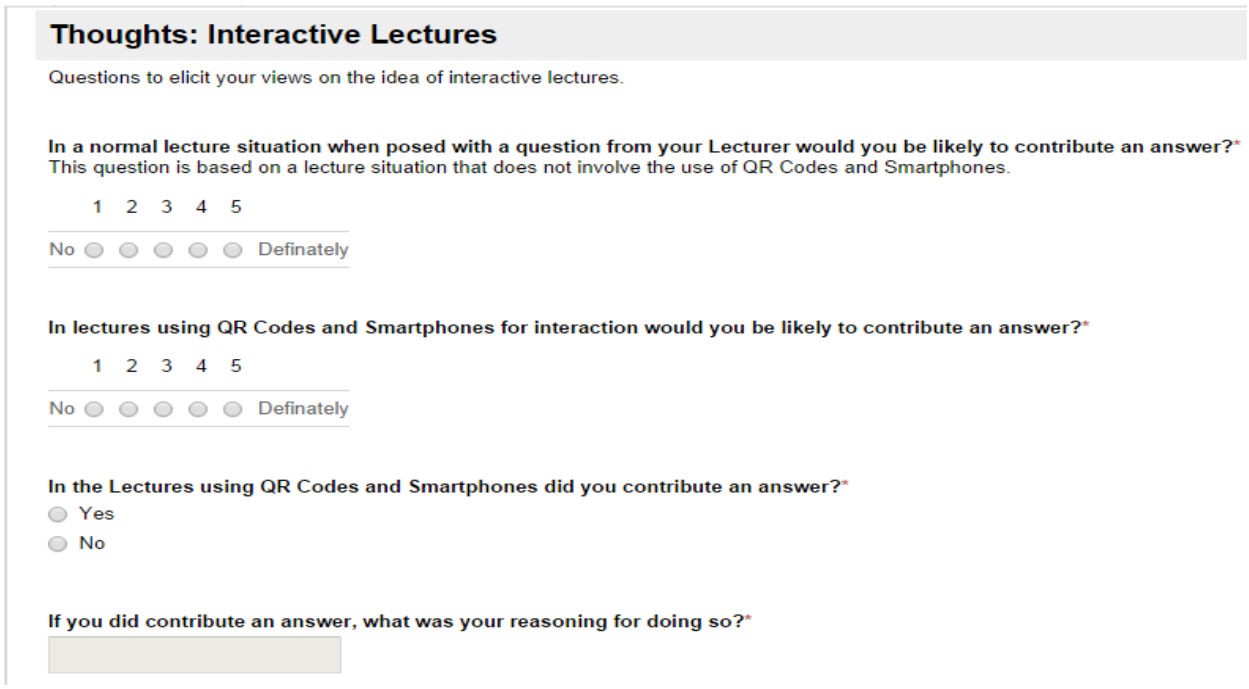


Figure 17. Survey Section 2

Figure 18 shows the form that is presented to the lecturer to help them build their question. The form collects the question, and answers and when submitted starts the automated process of creating the data repository and QR codes required.

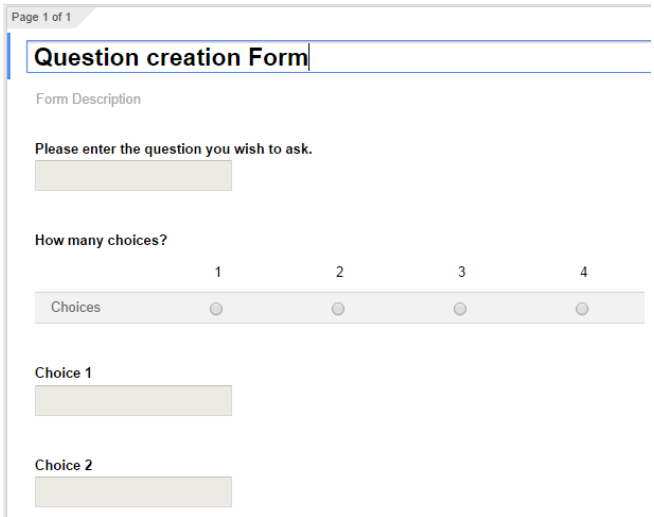


Figure 17. Form used to collate data required to build Q&A

Figures 3 and 4 help to illustrate how the automated process of creating the QR code is achieved by combining the data in the various cells and using Google's chart API to automatically generate the QR code.

This was not an inconsiderable amount of work to develop but as can be the problem with the reliance on a third party API based system Google has revised their API's requiring that the system undergo another rewrite to comply with Google's new spreadsheet, forms and graph generation tools. The new version of the system will be developed as a Google web based app. The upside to this is the fact that the user interfaces can be developed using standard HTML5 and CSS, thus more control over the look and feel of the system can be gained.

VI. STUDENT FEEDBACK

The students from each cohort were surveyed to ascertain some form of feedback to gage their thoughts on the system and pedagogical theory underpinning the system.

Selected results from the survey are presented within this section with accompanying thoughts and rational. The survey was split into three main sections; section 1 asking about hardware and software, section 2 asking about views on interactive lectures and section 3 asking for suggestions with regard to improvements to the system.

Figure 17 gives a flavor of the questions asked in section 2 of the survey. This section of the survey was used to try to build a picture of the students thoughts on interactive lectures and how likely they would be to interact with the lecturer under normal circumstances, i.e., answering questions by raising their hand and how likely they would be to answer a question using their smartphone.

The first question asked to the students in this section was "In a normal lecture situation when posed with a question from your Lecturer would you be likely to contribute an answer?". From the research outlined in Section II the expected answer was that few students if any would be tempted to offer an answer.

The chart in Figure 19 shows the result of the answer to the question. The question made use of a Likert scale rather than a straight yes or no to try and gage how adamant the students' answer would be.

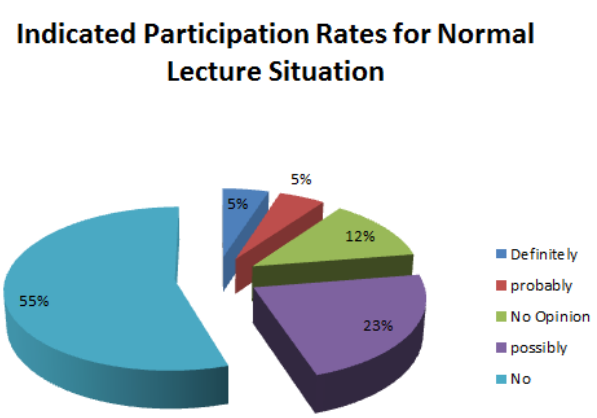


Figure 18. Intention to offer an answer in normal lecture situation

The percentages shown in Figure 19 indicate that the hypothesis expounded in Section II that students are reluctant to answer questions during lectures appears to be justified. Although 55% of respondents indicated that they would defiantly not answer a question the remaining 45% contained an interesting split. The two interesting percentages are the 23% who may possibly contribute an answer and the 12% who seemed ambivalent. It would be interesting to find out what would make the 23% possibles become probables or even definites.

The second question asked to the students in this section was "In lectures using QR Codes and Smartphones for interaction would you be likely to contribute an answer?" From the research outlined in Section II the expected answer was that almost all students would be tempted to contribute an answer.

The chart in Figure 20 shows the result of the answer to the question. Again the question made use of a Likert scale rather than a straight yes or no to try and gage how adamant the students' answer would be.

As can be seen from the results almost 90% of the students surveyed indicated that they would offer an answer. This was close to what was expected based on the research undertake for Section II. Interesting to note the 8% of respondents who had no opinion; further work will be required to ascertain why students opt for "no opinion".

The third question asked to the students in this section was "In the Lectures using QR Codes and Smartphones did you contribute an answer?" This question, on the surface, may seem a strange question to ask but it was an attempt to try to judge the accuracy of the scanned QR code data from the actual lectures.

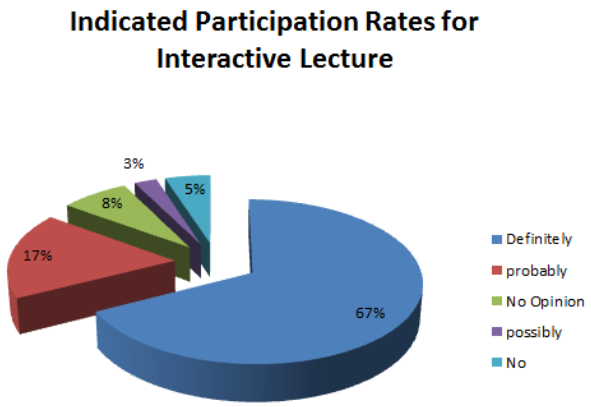


Figure 19. Suggested participation rates for interactive lectures using QR codes and Smartphones

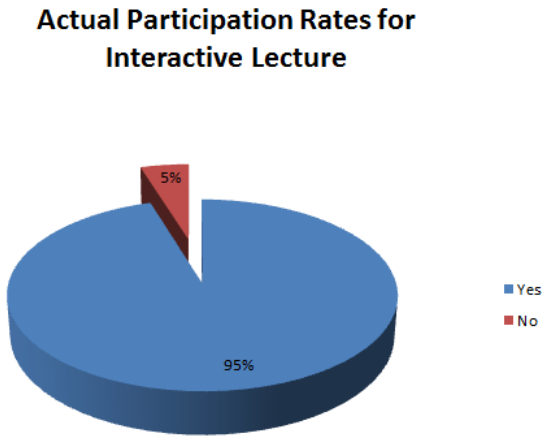


Figure 20. Participation rates for Interactive Lectures

Thankfully, the result to this question was an almost perfect match to the actual data gathered from the interactive lectures and had been the expect result.

Again, reflecting on the literature reviewed for Section II the author's interested was piqued as to why 95% of respondents felt that using the QR code system they would offer an answer. The obvious candidate was anonymity as cited by [14].

The fourth question asked to the students in this section was "If you did contribute an answer, what was your reasoning for doing so?" This question required the respondent to give a text based short answer. This was preferred to using any list based technique in order not to elude to answer of anonymity.

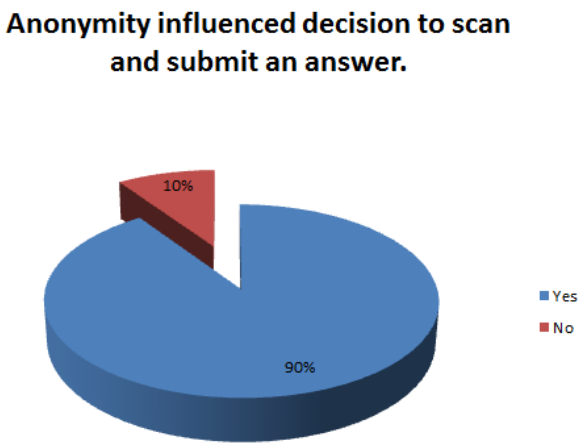


Figure 21. Anonymity plays a part on interaction

After sifting through and collating the responses to this question it became clear that hypothesis of anonymity was indeed a realistic prospect. Further questioning to gain a deeper insight into the students' thought process will be required in order to better understand the pedagogical issues underlying student contribution during lectures.

This section of the questionnaire more or less fitted with the ideas postulated in Section II, however, some of the results were not as clear cut as expected.

In the final section of the questionnaire, suggestions for improvements were sought to enhance the system and processes involved with using the system. One recurring improvement was to have more varied charts for displaying the results. This indicates that the students were paying a level of attention to the results more than superficial. It would be interesting to discover what the students think they can elicit from displaying the results as different charts and which chart types they would like to see and why.

VII. Conclusion and Future Work

Overall, the interactive lectures trialed with various cohorts have been generally well received by the students and the staff contributing to positive interaction between the lecturer and the students.

Students appeared to enjoy the break in the lecture and the feedback and discussion generated by the visual charting of their responses. It also created a focus point for the students to reflect on their understanding of the material taught and to apply that understanding. By the same token, it proved to be beneficial to the lecturer indicating the level of understanding of the delivered material to the students.

The anonymity of the whole process was cited by a number of students as positive and this has been borne out by the survey. The students gave the impression that they were comfortable with the fact that they could answer the questions freely, getting them wrong and not feeling awkward in front of their peers.

With regard to performance, this prototype system works well, producing the column chart of responses in near real time. Chart production will be expanded in line with suggestions made in the student survey to include line, pie and scatter charts. Advantages this system offers is the fact it is free, flexible, easily tailored to suit the lecturer's needs and platform independent.

A further avenue for investigation will be the correct utilization and positioning of the interactive lectures within the overall module lecture delivery schedule. Over or under use will have an impact on their effectiveness.

Further investigation will be made with regard to the sizing, positioning and visibility of QR Codes from projected and paper based slides. The rapidly changing hardware of smartphones and ever improving cameras suggests that the current inability of the majority of phones to scan a QR code from a projected slide may be at a demise within a couple of phone generations.

The use of the technique within the tutorial/seminar setting to encourage more debate on theoretical and social subjects is path that will be followed.

Further investigation will be undertaken into the relative pros and cons of storing complex responses in the spreadsheet, as evidenced in Figure 10, and simplistic responses in the form A, B, C, etc. The outcome of this investigation will have an impact on the future development of the software.

A significant proportion of future work will be involved in redeveloping the system to utilize HTML5 and CSS for the user interface to the software to allow cross discipline use. The script code will continue to be hidden from the user allowing them to concentrate on the development of their question bank and will be revised to comply with Google's criteria for use.

The project is ongoing and the positive feedback received from the students indicates that it is a worthwhile pursuit for both the lecturer and the students.

References

[1] R. Law, "Using Quick Response Codes For Student Interaction During Lectures," in ICCGI 2013, The Eighth International Multi-Conference on Computing in the Global Information Technology, 2013, pp. 29–33.

[2] Kantar Worldpanel, The smartest way to communicate: over half of the GB population owns a smartphone. Available: http://goo.gl/Z1n7u. [retrieved: July , 2013].

[3] Deloitte, "Media Consumer Survey 2014 The Digital Divide," 2014. Available http://www.deloitte.co.uk/mediaconsumer/ [retrieved September 2014].

[4] EUSA, Campus app now available for download. Available: http://www.eusa.ed.ac.uk/news/article/6001/290/. [retrieved: July , 2013].

[5] A. Ramsden, "The use of QR codes in Education: A getting started guide for academics," Working Article, University of Bath, 2008.

[6] R.Law, Using Quick Response Codes for Student Interaction During Lectures, ITiCSE, ACM 978-1-4503-1246, 2012.

[7] P. Gannon-Leary, C. Turnock, and M. McCarthy, "RECAP series article 15," 2007.

[8] Lowery, Roger C. "Teaching and learning with interactive student response systems: A comparison of commercial products in the higher-education market," In annual meeting of the Southwestern Social Science Association, New Orleans, LA. 2005.

[9] Murphy, Roger, and Namrata Sharma. "What don't we know about interactive lectures," In Seminar. net–International Journal of Media, Technology and Lifelong Learning, vol. 6, no. 1, pp. 111-119. 2010.

[10] D. J. Nicol and J. T. Boyle, "Peer instruction versus class-wide discussion in large classes: A comparison of two interaction methods in the wired classroom," Studies in Higher Education, vol. 28, no. 4, pp. 457–473, 2003.

[11] Angelo, Thomas A., and K. Patricia Cross. "Classroom assessment techniques: A handbook for faculty," Ann Arbor, MI: National Center for Research to Improve Postsecondary Teaching and Learning, 1993.

[12] R. J. Dufresne, W. J. Gerace, J. P. Mestre, and W. J. Leonard, "ASK-IT/A2L: Assessing student knowlede with instructional technology," arXiv preprint physics/0508144, 2005.

[13] C. H. Crouch and E. Mazur, "Peer instruction: Ten years of experience and results," American Journal of Physics, vol. 69, no. 9, pp. 970–977, 2001.

[14] V. Simpson and M. Oliver, "Electronic voting systems for lectures then and now: A comparison of research and practice," Australasian Journal of Educational Technology, vol. 23, no. 2, 2007.

[15] S.A Saravani and J.F. Clayton, "A conceptual model for the educational deployment of QR codes. In Same places, different spaces," Proceedings ascilite Auckland, 2009.

[16] V. Yfantis, P. Kalagiakos, C. Kouloumperi, and P. Karampelas, "Quick response codes in E-learning," Education and e-Learning Innovations (ICEELI), International Conference, 2012.

[17] C.Y. Law, and W.W.S. So, "QR codes in education," Journal of Educational Technology Development and Exchange, vol. 3, no. 1, 2010, pp. 85-100.

[18] E. Bojinova and J. Oigara, "Teaching and learning with clickers: Are clickers good for students," Interdisciplinary Journal of E-Learning and Learning Objects, vol. 7, 2011, pp. 169-183.

[19] A.P. Fagen, C.H. Crouch, and E. Mazur, "Peer instruction: Results from a range of classrooms," Physics Teacher, vol. 40, no. 4, 2002, pp. 206-209.

[20] S.R. Jones, "Was there a Hawthorne effect?," American Journal of Sociology, 1992, pp. 451-468.

[21] J. Devon and R. Law, "Action Research Project Looking At Providing An Environment To Improve Student Engagement Using Audience Response System Considerations And Smartphone Technology," ICERI2013 Proceedings, pp. 969–978, 2013.

[22] I. D. Beatty, W. J. Gerace, W. J. Leonard, and R. J. Dufresne, "Designing effective questions for classroom response system teaching," American Journal of Physics, vol. 74, no. 1, pp. 31–39, 2006.

[23] S. Sandhu, T. Afifi, and F. Amara, "Theories and practical steps for delivering effective lectures," J Community Med Health Education, vol. 2, no. 158, pp. 2161–0711, 2012.

Ecologies of Spaces for Enjoyable Interactions

Alma L. Culén and Rune B. Rosseland
Department of Informatics, Group for Design of Information Systems
University of Oslo
Oslo, Norway
almira@ifi.uio.no, runebro@ifi.uio.no

***Abstract*— In this paper, interplay of diverse factors related to the space where public interactions take place is discussed. The technology itself, the physical space, activities and social interactions around them are all important for the user experience, particularly so when enjoyable interactions are in focus. We call this plethora of factors and relations between them the ecology of an interactive space. Concepts such as visual immediacy, impetus and impedance, related to exhibits access and entry points are introduced in order to discuss engagement with installations, but also the space and social interactions. We illustrate this using the installation for enjoyment that we designed, implemented and subsequently evaluated in three public settings. Drawing on our findings from the experience with the exhibit, as well as conceptual and practical research related to interactive exhibits, we conclude that the concept of ecologies of spaces is useful for deeper understanding and design of public interactions.**

Keywords*— *interactive installations; play; public space interaction; user experience; Kinect; ecology.

I. INTRODUCTION

Public spaces are increasingly also interactive spaces [1]. Small portable devices such as i-beacons, large and small interactive screens, diverse sensors, mobile and tangible devices enable design of interactive zones in public spaces relatively easily and inexpensively. Consequently, interactivity in public space is becoming ubiquitous. Some interactive installations have a specific functional purpose, e.g., touch based information boards, check inn points, etc. Others are more geared towards inspiration, reflection, art or entertainment. Interactions in this latter group are often designed for specific places, e.g., museums (Fig. 1), bookstores, cafés (Fig. 2) or galleries, often with no other purpose than to provide enjoyable experiences.

Figure 1. The New Children's museum in San Diego offers an interactive DJ table with graphic display. Photo: Culén.

Figure 2. Funky Forest exhibit at Moomah café in New York offers young visitors an interactive experience of a forest ecosystem, photo from [2].

These places differ in how they engage their audiences. The galleries often focus on sensory experiences and wow factors, museums on new learning and knowledge constructing opportunities, while cafés and bookstores may extend a more commercial variety of offerings.

An interesting new arena for enjoyable interactions is a workspace. There is increasing evidence that mood and creativity are deeply intertwined, see an analysis of a 25 years long study on this relationship in [3]. Supporting good work environment seems to facilitate creative processes and collaboration [4][5]. Thus, many companies, e.g., Google [6], are trying the lighten the mood of their employees using playful and enjoyable interactions, and capitalize on heightened employees' creativity. A likely future trend is expansion beyond museums, galleries, workspaces etc. into hospital waiting rooms, centers for elderly, airports, elevators and all other less grand public places where people may benefit from lightening up.

Thus, as multi-sensorial, playful interactions enter the public sphere, it makes sense to look into what kinds of public space configurations are suitable for interactive installations that provide enjoyable, open-ended experiences and co-experiences. In [1], Rosseland, Berge and Culén discuss how user experiences with an interactive installation were influenced by the contextual setting of the installation. The installation was designed to provide multiple, co-located users with an enjoyable audio-visual experience in response to gestures and bodily movements, and was tested publicly in two different settings: a university library and a Mini Makers Faire at a science museum. The success of the installation was measured in terms of engagement time.

This paper uses the same case as Rosseland, Berge and Culén's paper [1]. The novel contribution presented in this extended paper is that of a conceptual framework, which can be used to discuss the design and evaluation of enjoyable interactions in public spaces. To this end, we propose the concept '*ecology of interactive spaces'*. The term is used to denote the complexity of material and social relationships that exist in any public setting with interactive technology. The paper aims to start a dialogue that examines the role of the material space and social practices when designing for enjoyable and social public space interactions.

The structure of the paper is as follows: in Section II, we provide some background needed to understand how our framework relates to previous research in the field. We also discuss the work related to enjoyment, pleasure, play and games. Entry and access points to installations are briefly discussed. In Section III, we present our concept of the ecology of interactive spaces, with physical space attributes, technologies, activities, people and values as central components of the proposed concept. Section IV provides details from our case study of enjoyable interaction. In Section V, the concept of ecology is applied to the described case. Discussion and conclusion follow in Sections VI and VII, respectively.

II. THE BACKGROUND

The concept of the ecology of interactive spaces was conceived under the influence of Nardi and O'Day's work [7], our previous work, and work by others, most notably [8]–[13]. From [8], we take with us the importance of designing beyond products and including activity spaces around interactive products into consideration. From [9][10] we learn to pay attention to entry and access points and that tangible interactions can enrich interactive spaces. Papers [11][12] address issues of understanding experiences and defining a concept of co-experience. Co-experiences are defined as user experiences through social interactions, and are central for public space interactions. From [13], we understand how even simple chaining of displays into different shapes influences by-passers and users, single and in groups Fig. 3.

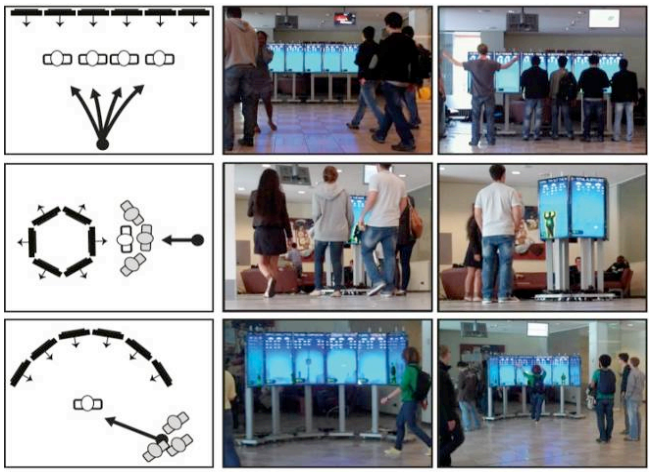


Figure 3. Diverse screen configurations lead to different behaviors, [13].

A. Enjoyment: Pleasure, Fun and Play

In order to design installations for public spaces that build on fun, pleasure and play, a co-located presence of other people, both acquaintances and total strangers, becomes an important factor for engagement. HCI research in this area may draw on social science research within the museum field such as [14], as well as on HCI research related to shareability [10] and enjoyment, e.g., Blythe and Hassenzahl's work [15], where the semantics of pleasure and fun is discussed. We consider the enjoyment first.

Enjoyment can be thought of as an experience fleeting somewhere between distraction and absorption, where, on one end, fun represents distraction, and pleasure represent the absorption side of the scale. In short, fun is described as the counterpart to seriousness. As a distraction, it represents a spontaneous escape from the tasks and worries of everyday life. The self, the hedonic 'be-goals' of UX, does not matter in this short-lived break from reality, but fun still satisfies an important psychological need.

Pleasure is found on the opposite end of the enjoyment scale, taking on the role of absorption. It represents a deeper, longer lasting, more meaningful experience. Here, the connection to people's inner self is made through immersion and devotion to an activity. Elements of challenge, progression, and demand for absolute concentration can be present, and thereby overlaps with Csikszentmihaly's concept of flow; see [16] and [17].

Play is another fuzzy term to corner, as illustrated quite well by Sutton-Smith who has dedicated a whole book to this topic: "We all play occasionally, and we all know what playing feels like. But when it comes to making theoretical statements about what play is, we fall into silliness. There is little agreement among us, and much ambiguity", [17, p. 1].

Although the term play represents a myriad of experiences, it has been broadly described as a "*free movement within a more rigid structure*" [18].

Some of the most influential work on play is done by the French sociologist Caillois. He divides play into four forms and two types of play [19]. The four forms of play are competition, chance, simulation and vertigo, and the two types of play are free play and formal play [20].

Playful behaviour is described as an oscillation between exploration and engagement [21]. Playful behaviour starts with exploration, and play occurs when the unfamiliar becomes familiar [22]. When the familiar gets boring, the focus returns to exploration. In this context, the goal of exploring is best described by contemplating "what can this object do?" The goal of play, though, is related to the question "what can I do with this object?"

Games, as opposed to play, have a structure, as well as an objective to reach. It has been shown that play and playful design, including games and gamification, have a large effect on users' motivation and engagement [23].

B. Entry and Access Points Re-visited

In [10], Hornecker, Marshall and Rogers discuss the concept of shareability. Shareability is defined as a design principle that refers to how a system, interface, or device engages a group of co-located and co-present people with the

same content or the same object. The authors propose entry and access points as the aspects of shearability. These concepts are related to two levels of user engagement: the user needs to 1) be motivated to use the installation 2) know how to use it.

Entry points denote "design characteristics that invite people into engagement with a group activity and entice them to interact. Access points denote characteristics that enable the user to actually interact and join a group's activity" [10, p. 3].

However, rather than distinguishing different components of entry and access points as described in [10], e.g., honeypot effect, minimal barriers, perceptual access and others, we propose a somewhat different approach, starting from the work of Gardner. According to Gardner's theory of human intelligences, [24], people posses unique combinations of visual–spatial, verbal–linguistic, logical–mathematical, bodily–kinaesthetic, musical, interpersonal and intrapersonal intelligences, which orchestrate our understanding of the world and define our actions in it. So, naturally, these will have effect on how each person enters and accesses the sharable space. When entering a shared interactive space, the visual-spatial input would be of highest significance [24], as an overall assessment of the space, and activities in it. However, in the case of the installation that we describe in Section IV, it is easy to see that musical, body–kinaesthetic, intra and interpersonal intelligences strongly relate to the experience of the installation.

In discussing entry and access points, we propose the use of concepts of *immediacy*, *impetus* and *impedance*, as well as *fluidity of sharing* introduced in [10, p. 3,9]. Visual immediacy, see [25][26], is proposed as a characteristic of visual–spatial intelligence, to help reason around interactive space initially. Immediacy gives the first impression of the space in terms of safety and appeal, as well as the initial understanding of the activities taking place within the interaction area. Impetus gives a nudge to engage in activities and impedance represents barriers, resistance or hindrances to enter the interactive space. Thus, the honeypot effect, for example, may be considered as a factor that gives impetus to majority of people to enter an interactive space. Lastly, in contrast to affordance, a design characteristic referring to just those action possibilities that are readily perceivable by an actor, impedance is a characteristic related to action possibilities that are difficult to perceive due to the existence of barriers, hindrances or resistances.

Given that interactions in public spaces are strongly influenced by other co-located and co-present people, we need to consider both experiences and co-experiences. Fluidity of sharing captures how easy it is for people to engage in joint interaction and creation of co-experiences.

C. Interactive spaces

The spaces that come to mind in relation to public space interactive installations that have been already researched to some extent are museums, libraries and workspaces.

Museums are increasingly involved in providing digitally responsive exhibits, as part of their strategy to attract visitors. These exhibits, as Heath and Lehn point out in [14], often involve diverse displays enabling either individuals or groups to promote thinking and discussion around material presented in the museum, thus providing additional possibilities for learning and understanding. Enabling learning is often the main goal of museums. Consequently, museums' approach to evaluation of their interactive exhibits favours standard methods used in the museum field: focus groups, interviews and questionnaires, focusing on the learning outcomes. Thus, the experiences and co-experiences during the interactions are usually not the focus when applying the above methods. In addition, although people often come to the museum in groups, very few exhibits are explicitely designed for co-experiences [27], [28], or, as mentioned, for pleasure.

Similarly, workspaces are also increasingly focusing on playful and enjoyable experiences at work, reasoning that such experiences may increase satisfaction with the work place, increase productivity and provide an easy entry point for people to meet each other [29].

In their paper [8], Kaptelinin and Bannon propose for the field of interaction design to move beyond design of products and into design of technology-enhanced activity spaces. The article presents three related arguments. The first one has to do with the fact that technological development so far has provided more support towards extrinsic rather than intrinsic human practices. This concern is related to opening the space for practices that are initiated by users. This leads to the second argument, the one concerning the 'ecological turn', nudging the field of interaction design to develop methods that allow intrinsic practice transformations. This is raising some of the same concerns as those in [30], employing the 'semantic turn', focusing on meaning of technological design interventions in the real world. The third argument is a direct invitation for the field of interaction design to expand to include creating technology-enhanced activity spaces.

The word ecology and related concepts such as habitat, species and environment, has been used in HCI for a while. After information ecology was introduced in [7], it became a common metaphor for describing complex relations between local environments, technology and people.

III. The Ecology of Interactive Spaces

Taking up the challenge presented by [8] to engage with design of interactive spaces, we choose to use the concept of ecology.

When considering *ecologies of spaces* for public interaction, we propose five main components:

- *Space* - including its properties such as materials, spatial layout, acoustics, light, and aesthetics, as well as the larger space of which it is part of, e.g., a hospital, a museum etc.
- *Technology* - the installation itself, but also technology that is either part of the environment or brought in by people, e.g., smart phones, sensors in the room.

- *People* – who use and inhabit the space; decision makers/owners of the space; designers and other stakeholders
- *Activities* - that are conducted in the space, using the installation, the space and other people to create experiences and co-experiences.
- *Values* - the explicit and implicit values, norms, rules that co-exists within a given location.

In terms of design for public installations, each of these five main components of the ecology opens for new design opportunities. Being aware of all five, and that they need to work well together, may help designers to design better installations for enjoyment in public spaces. For example, the above mentioned paper by Hornecker, Marshall and Rogers [10] introduces shareability as a design principle for design of interfaces that engage co-located, co-present users in shared interactions. More concepts such as that of shareability are needed in order to generate a set of design principles covering all aspects of the ecology.

In terms of evaluation, we do not know of any previous work concerned with evaluation of an installation as a whole, including all components of the ecology. What we find in the literature are frameworks for evaluation of diverse aspects of interactive installations, e.g., [20], [28]. Furthermore, in [28], it is pointed out that there are no frameworks for evaluating social interactions and co-experiences in museums, and we find this to be true for other public spaces as well. Yet, public interactive spaces are designed for technology supported social interactions, see [31] as an example. We are unaware of any framework for evaluating installations as a whole. Our previous paper also focuses only on the comparison of spaces, and only based on the length of time that participants engaged with the installation. In this paper, we take initial steps towards a broader and more systematic view when evaluating the installation, a view supported by the introduced concept of the ecology of interactive spaces.

We now present details of how the installation was designed and tested in the lab, as well as what we learned from the lab testing, much along the lines presented in [1]. In Section V, we relate introduced concepts, not only as they pertain to the interface, but also the space, activities, people and the social component of the ecology.

IV. The Case of an Installation for Enjoyment

We now describe the basic set up of the installation. The set up description is followed by results related to the concepts of fun, play and pleasure (as discussed in the Section II) from user evaluation of the installation in the lab.

The installation was designed to give a pleasurable experience. It did not solve a problem, nor did it aspire to help people reach meaningful life-goals. Its purpose was, ultimately, to be a research tool and give us as designers and researchers the opportunity to observe, evaluate and learn something about user enjoyment and behaviour in both public (library and museum) and a more private context of the lab. The installation was designed for pleasure, to be enjoyed individually, or with others, familiar or total strangers.

The installation consisted of:

- A long and narrow table placed by a wall.
- Two Kinect sensors mounted on the table on top of each other.
- One Shake 'n' Sense device [32], fastened to one of the sensors to eliminate interference.
- A wall-mounted screen, either a flat screen TV or a canvas lit by a projector.
- Two amplified speakers placed on the table on each side of the screen.
- Two Mac laptops placed outside of the installation area, one running the audio and the other the visual system.

In all locations, the installation was exhibited in the setup as shown diagrammatically in Fig. 4, and in the actual space in Fig. 5. Each of the locations had, at least, an area of four by four meters in front of the Kinect sensors. The installation consisted of two completely separate systems, one controlling the audio and one controlling the visual display. The systems were tuned to work together and appeared for the user as a single installation.

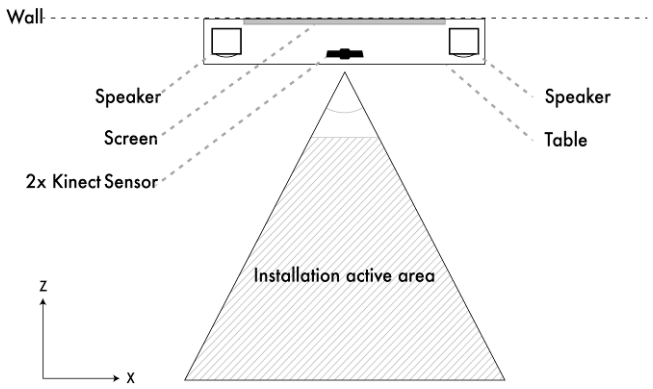


Figure 4. Diagram representing the physical setup of the installation.

When a user, or several users, walked into the range of the sensors, the system automatically identified them, and started tracking their movement and playing music.

Figure 5. Actual setup of the installation in the museum during the Maker Faire, with interaction area marked on the floor. Photo: Berge.

The main way a user could interact with the installation was by extending an arm away from their body. More specifically, a horizontal hand movement away from the chest would trigger the system and start the calibration processes, engaging both audio and visuals.

Before testing the installation in a public setting, we conducted lab evaluations to get some feedback on how the installation was perceived by users who had never seen it before, as well as identify opportunities for improvement. The lab, as a space, is a large room, with a single door access, not allowing any view of activities from the outside when the door is closed. Thus, it was ideal for private and personal exploration of interaction, allowing it to unfold without having to consider how the process looks like for others. We invited people in pairs in order to see how having to share the same interaction space affected the participants. All kinds of explorations were welcome, both on an individual level and in pairs.

The ten participants were all either current or former students at the Department of Informatics. The sessions were videotaped, and the participants were interviewed immediately after finishing the test of the installation.

Through the analysis of our evaluation sessions in the private lab context, we located many statements indicating an enjoyable experience. In this section, we want to look at aspects of the participants experiences related to the concepts the terms enjoyment, as discussed in Section II.

A. Fun

The installation in itself was described by most as 'fun'. Blythe and Hassenzahl defined fun as a short-lived distraction from everyday life [33], coinciding well with the way the word is used in describing the experience by the participants. But what exactly was fun about the installation? The participants' answers point first and foremost to the exploration of the installation and its functionality, and then secondly, to the immediate responses the installation gave to movement, and the sensory aesthetic experiences these movements resulted in.

Pleasure was never mentioned directly by the participants, but several interviewed participants described an experience of 'flow' [34] when they were interacting with the exhibit, which can be linked to pleasure [15]. These experiences were described in terms of being 'lost', mesmerized, having a break from thinking and entering a relaxed 'kind of mode'. The majority of the participants agreed on this being an essential part of their experience. It is worth noting that some of the participants pointed out, both explicitly and implicitly, that this flow-like state disappeared over time as the participants ran out of elements of the installation to explore.

B. Play

Several of the participants described the installation and the experience as playful. Their descriptions indicated that they placed the experience more in line with the definition of free play, rather than structured play (game).

The playfulness that the installation enabled was deemed as very important, and the participants linked it strongly to the exploration and the open-endedness of experience, but also to the lack of control. The openness of the installation was described as an advantage, in the way that it encouraged interpretation and exploration. The lack of control was described as not important by one participant, as the point is not to steer something, but to play with the system and get responses from it, which resulted in a 'good feeling'. In relation to the concepts of goals, rules, and competitive elements of play, even the self-proclaimed 'competition-focused' participants acknowledged that those concepts were not the point of this installation.

In the playful behaviour there is an oscillation between exploration and play, where play is triggered by learning or discovery and exploration is triggered by boredom [21], [22], [35]. We found multiple instances of this in the way participants described their explorative behaviour, which strongly resembles the process of playful behaviour, emphasizing the strong relation between playing and exploring: *"It is just exploring, really. Until you feel you master (the installation) a bit, then it's really exciting and makes you want to continue. You never know if you have explored everything and that's positive, you never reach an end."*

C. Aesthetics

In terms of aesthetics, both the audio and the visuals were described as fascinating, atmospheric, different, beautiful and soothing. The participants thought the combination of the two fit well together and resulted in a coherent expression and created a good ambiance. It was also pointed out in a positive manner that the expression was kept to an abstract nature. That way it became easier to accept the audio-visual expression, in comparison to trying to depict or simulate something concrete.

D. Exploration

As stated earlier, exploration was the main activity that the installation was designed for, and the experiences during the exploration were deemed to be the most important, successful, aspect of the installation. Several of the participants expressed bluntly that exploration *is* the installation. The exploration was fuelled by the responses given by the installation and their abstract, mysterious, unknown nature. Or, to put in other words, the immediate responses to movement and actions, combined with lack of explanation, made the participants curious and eager to investigate. Their descriptions also highlighted one of the common characteristics of the human brain, namely the constant search for patterns and connections, which was described as an essential part of the process of exploring.

E. Discovery, learning and understanding

On some aspects of the experience, the participants were quite divided in their opinions. One of these aspects was the lack of explanation, or guidance, in the user interface of the installation. The majority of participants highlighted the absence of explanations as something positive. It was seen as a catalyst for, and a component of, exploration. However, some found it confusing, frustrating and incomprehensible.

One of the participants, who favoured minimal explanations, pointed out that an installation such as ours would not be suitable for people who are not interested in exploring.

The discovering and learning were described as closely related to exploration. For example, one participant described discovery as a direct result of the exploration.

The process of understanding was the challenging part of the installation. When exploration led to discoveries and understanding, the participants had a sense of progress and achievement, giving them motivation to continue to explore. However, the lack of 'new things' to discover and explore eventually led to boredom and loss of interest.

F. Progression

Progression was the aspect of enjoyable user experience that was originally overlooked in the design process, but which surfaced through the evaluation of the prototype as the most important missing aspect of the participants' experience. As mentioned earlier, Blythe and Hassenzahl link the concept of pleasure to the concept of flow, but they also argue that pleasure can in fact be thought of in terms of progression [15]. In retrospect, this actually comes across as self evident, when comparing our findings to the overlapping definitions of flow and pleasure, as a longer lasting, more meaningful and immersed experience devoted to an activity.

The participants wanted more depth to the experience. They wanted more to explore, with gradual increase in variation and difficulty. When they felt they had exhausted their possibilities for exploration, they became bored, and this coincided with the earlier mentioned loss of the state of flow.

G. Control

The second most sought after aspect was control. On this point, the participants of the prototype evaluation were close to unanimous. They expressed frustration over not getting the expected responses from the system, and this put limitations on what they could do. It prevented them from being creative and expressing themselves through the installation, both in terms of visual and audio expression, and this was considered to be of large importance for them. Some acknowledged that they attained a certain degree of control, but they expressed that the threshold for gaining this control should be much lower in order to make this aspect of the installation accessible to more people.

The lack of control linked very strongly to the absence of mastery, and on this point the feedback from one of the users was quite direct: *"*[The installation] *lacks possibility for mastery."* And another user on the same topic: *"I don't think I would master it more if I used it for another 20 minutes."*

The feedback we got from the participants brought forth the distinctions between the second and the third paradigm of HCI [36], and between usability and user experience. In our phenomenological approach, the focus was on enjoyable user experiences, and not so much on usability and ease of use. Also, the explorative and abstract nature of the installation meant that it was difficult to define specific usability criteria for it.

This is not to say that control was not a focus in our design, but the lack of precision in the tracking data from the Kinects, and our experiential focus led us to design a system that, we thought, did not need very specific and precise controls. Nevertheless, our findings clearly show that lack of control detracted from the experience.

The first public test of our installation, apart from one exhibit in the lab open to general audience, was at the Science Library at the University of Oslo, in two different locations. Subsequently, the installation was tested in a museum setting during the Makers Faire days. We now present our findings from these public exhibits of the installation and from the perspective of ecologies of interactive spaces.

V. Ecologies of Spaces for Enjoyable Interactions

A. Ecologies of space for enjoyable interactive installations in an academic library

The science library actively encourages students to develop and use different kinds of systems and technologies, in the library. We were invited to set up our installation in the foyer on the ground floor of the library building for three consecutive days. This provided a good opportunity to observe how people react to and interacted with the installation in the wild, in the realistic public setting where the exhibit could be a more permanent one.

1) The space

The Science library is a large three stories high brick building, with lots of open space just across the main entrance into the building, see Fig. 6.

Figure 6. The entrance area is used as a café, a stage, for sharing information etc. The screen used for the exhibit is marked. Photo: Juell.

The facade and doors on the ground floor on the entrance side are covered with floor-to-ceiling windows, providing ample daylight and allowing people from the outside to see what is going on inside. Fig. 7 shows the exact position of the installation during the first two days of observation. This is a location directly opposite the main entrance and part of the very open area, as shown in Fig. 6.

Figure 7. The installation at the open space, across from the main entrance into the building, giving a different perspective of the space. Photo: Culén.

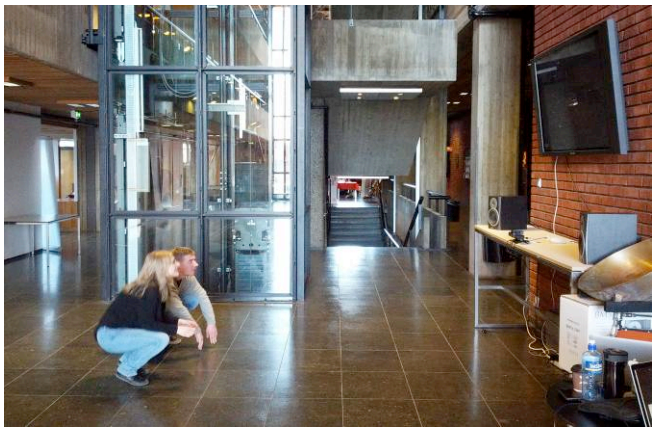


Figure 8. The same local environment, the library, with different exhibit space – less exposed to others. Photo: Berge.

On the third day, a more secluded area of the library was used, see Fig. 8. This was done, in part, to allow people to feel freer when exploring the installation.

2) The people

The people frequenting the library are mostly students, and some faculty and other staff. They are either alone or clustered in small groups of 2-5 people. On the third day set for our observations, there was an event for graduating high-school students on the ground floor. During the event, the area filled up with 70-100 pupils, but relatively few found the installation, and only 10 pupils actually engaged with the installation.

The predominant kinds of relationships between the people at the library include friends, classmates, strangers, the significant other, colleague, employee, faculty, student, and the occasional sibling / parent / spouse.

3) The activities

Typical activities in the ground floor of the library include eating, studying / working, socializing, flirting, waiting, and walking. Most people were preoccupied with their own activities, and were not paying much attention to what other people were doing. They were there because they had some business there, either going to or from a lecture or the library, or to kill time until the next activity on their schedule.

However, the activities in this space can change quite dramatically, with little or no warning for its unsuspecting regular visitors. A good example of such a change is the infusion of scores of high school seniors, which influenced the usual activity patterns. Thus, the open area on ground floor of the library is not entirely predictable space in terms of the activities that may take place there.

4) The technology

There is very little technology in the library foyer that is part of the normal inventory. There is the usual set up at the stage, see Fig. 6, with a projector and diverse points to connect laptops and mobiles in order to show presentations, movies and mediate discussions. Otherwise, there are just a few computer terminals and a few large screens, two of which were used for the installation.

The TVs on the wall were hanging there permanently, usually displaying information regarding activities in the library, and the visitors were quite used to their presence. Our installation introduced a table with two speakers at either end, and two Kinects mounted above each other in the middle of the table. This constituted a fairly unnoticeable change from the usual setting.

5) The values and norms

There are numerous rules, norms, and values about how to conduct oneself in a particular public space. The library is no exception. Here, though, most of the norms are implied, rules unwritten, and are only enforced to the extent that people feel bound by them. Prevalent values at the library can be described as: be cool, different, similar, attractive, helpful, friendly, tolerant, competent and capable, curious and inquisitive, do not disturb others, be quiet, apply yourself, but do not overdo it.

6) Findings from the library

During the three days period when interactions with the installation were observed, we have spent a total of 7 hours 49 minutes, distributed as follows: 2h on the first day, 2h 36m on the second, and 3h 13m on the third day. 52 interaction sessions were observed, some of them involving groups. The participants were mostly in the age group between 17-40 years old, with few older exceptions.

The granularity of time-registration was not fine enough to draw any certain conclusions regarding time spent with the installation, other than that hardly anyone spent more than three minutes. However, when comparing the average time spent on group vs. individual interactions, we see that groups spent more than twice as much time then individuals interacting with the installation alone (1.2 minutes vs. 0.5 minutes in average).

Perhaps in contrast to the usual absence of music in the library, people soon learned that whenever the music started, there were people interacting with the installation. This allowed them to look up whenever the installation was in use, thereby slowly building an understanding of how it worked. This also allowed them to build both curiosity and courage to engage with the installation. We saw several

examples of people coming up to investigate after having observed others interacting with it for a while. There were also examples of single persons and groups of people who were hanging around in the background, queuing when others were interacting with the installation. As soon as the people using the installation left, they would walk up and give it a try. This worked like a honeypot effect, a positive feedback loop, where use attracted attention and instigated more use. However, the installation was unable to keep people's interest for more than a minute or two, which meant that there would have to be a constant stream of people to keep the installation in continuous use. When the installation was allowed to go into standby mode, people quickly returned their attention to whatever they were otherwise doing.

In terms of the level of engagement, people who explored the installation together with others seemed to get more out of it than those who were alone. They would talk to each other and explore cooperatively, discovering more functionality. There were also several examples of people who had been interacting with the installation earlier came back with friends.

Verbal reactions were usually immediate and short, perhaps also because the observers were hidden, looking just like everyone else, so people were more or less talking either to themselves or to their friends:

"Awesome! Motion sensor, cool!" – Man X
"Shit! Wow!" – Girl A
"Very cool!" – Man Y
"Pretty cool!" – Man Z

There were also more reflective statements:

"It responds to my movement." – Man W

After exploring for a minute, one man, of about 60 years old, exclaimed: *"One could stand here all day, fooling around!"* – Man P.

Many participants explicitly mentioned the word cool. Coolness of technological objects may be an important factor for their acceptance, as well as a design goal, see [37].

Figure 9. The Maker Faire takes place in diverse locations within the Norwegian Science Museum. Photo: Juell.

B. *Ecologies of space for enjoyable interactive installations science Museum / Maker Faire*

The second public test took place at the Norwegian Science Museum, at the Mini Maker Faire in Oslo, Fig. 9.

In the Museum, we observed users just one day, for 42 minutes. During the observation time, 33 individuals interacted with the system, some alone and others in groups, see Fig. 10 and Fig. 11. The age span of participants was from ca. 1 year old to elderly well over 70.

1) *The space*

The Norwegian Science Museum is a large museum, receiving about 250 000 visitors per year. The museum offers many different exhibits in exhibit rooms and in open areas. Similar to the library, the main entrance leads into a large open area with a reception directly in front of the entrance, a cafeteria to the right, and an open area to the left leading into diverse permanent exhibit areas.

During the Oslo Mini Maker Faire, this open area was the main exhibition area, and was filled with tables and stands with a plethora of different projects and technologies on display for visitors to explore.

The area that our installation was set up in was within the area for permanent exhibits, in a D-shaped room, see Fig. 10.

Figure 10. Interacting with the system at the Norwegian Science Museum. Photo: Culén.

We used a projector to project the graphics onto the straight wall, and mounted the Kinects on a table directly below the projection area. We used black tape on the floor to delimit a triangular interaction area corresponding to the horizontal field of view of the Kinects, in order to make it more comprehensible for the visitors where they needed to stand to interact with the installation. Furthermore, in anticipation of visitors arriving in small groups of friends or families, we placed small sitting cubes along the sides of the interaction area where onlookers could sit down and wait while their friends / children / grandchildren explored the installation.

2) *The people*

Typical visitors at the Norwegian Science Museum are families with children in the pre-, primary-, and middle school ages, as well as classes from schools around the city.

However, during the Oslo Mini Maker Faire, which we were a part of, the visitors included a wider mix of people. This was to a great advantage for the evaluation of the installation, as there were more people with expectations to be surprised and engaged. There were university students, researchers, volunteers, model train enthusiasts, people dressed in medieval and science-fiction costumes, and makers and tinkerers of all ages.

The predominant relationships among the people in this space were family, friends, or strangers. There were some colleagues, classmates, neighbours, and other acquaintances, but visitors mostly arrived with family and friends, in small groups of 2 to 6 people. To them, practically everyone else was a stranger.

1) The activities

People come to the museum to experience, learn, and enjoy themselves. But within the Maker Faire context, there is much more focus on the social, there is more noise, more exploration and interaction, and more of hands-on experiences, when compared to permanent exhibits. In part, the focus in on innovation and mastery of do-it-yourself variety.

Regarding the installation, before arriving to the exhibit area, there was not much that deterred attention. If one chose to follow the way to the opening of the room with the installation, one was usually drawn in to engage with the exhibit, or, to sit and observe others engage with it.

2) The technology

The museum on the day of the Faire had many new makers projects utilizing diverse types of technologies. Permanent installations, mostly around technology, were also present.

Our installation utilized different display than in the library. Rather than using a TV display, we used the projector and the white wall in this space. The remaining technological components of the installation were not changed. Here, as in the library, people had their own mobile phones, cameras, headphones and other small mobile devices.

3) The values and norms

In contrast to the library, there are fewer norms to follow, in particular during the Mini Makers Faire days. The values are different and oriented towards innovativeness, creativity, mastery, joy, play, experience and socialization.

4) Findings from the Maker Faire

In terms of the level of engagement, people have been engaged with the exhibit for a longer time than in the library, even when interacting alone, as was the case of a young boy, ca. 5 years of age, who spent 5 minutes exploring. For groups, spending 4-5 minutes was common. The group members would talk to each other and explore cooperatively, indicating, in line with the library observations, that it was fun to share. Here too, there were examples of people who came back with friends or family.

Verbal reactions were similar to those at the library:

"Cool!" – Boy X

"This was fun!" – Boy Y

"This was really fun!" – Boy Z

"Do I influence the music? … Oh, I see, I do!" – Lady A.

Comparing the engagement with that in the library, we found also that people had a lot lower threshold to join someone who was already exploring the system. It was interesting to observe that intergenerational interactions were not uncommon. Children would freely join adults whom they did not know. Very young children also tried the installation, as shown in Fig. 11.

We are now in a position to present overall findings from our exploration of this installation, focusing primarily on the space, and measuring engagement in terms of the length of engagement.

Figure 11. The installation engaged all age groups, also frequently the children shared the space with adults, even when they did not know them. The child on the right hand side of the photo, was under two years old. Photos: Culén.

C. *Overall findings related to differences in enjoyment between the library and the Maker Faire*

Most of our interviewees from prototype testing sessions in the lab readily admitted that they would restrain their involvement with the installation in a public setting, if they would be willing to interact with it at all. The most central reason they gave for this was the fear of breaking social rules and norms, and of "behaving like an idiot", as one participant put it. It was their fear of being perceived by others as doing something people do not normally do in public that would keep them from getting too involved. There were also comments to the opposite effect, indicating that breaking social rules and norms can be liberating and empowering. However, the prevailing notion was that social rules and norms would have a dampening effect on people's level of involvement with such an installation in public settings. Some participants consistently underlined that they would be less likely to interact with the installation on the street or in the shopping center, than a destination like a museum or a gallery. One of the participants expressed this as follows: *"It would be a lot more socially acceptable in a museum to interact with it. I would say my experience would have been much better in a context like that. If the installation were set up in Karl Johan* [note: central shopping street in Oslo] *I wouldn't have stopped to check it out, also because I'm going somewhere"*.

This concern seemed particularly evident at the library. The openness of the location and the number of people in the surrounding area seemed to make people self-conscious and vulnerable when they triggered the installation, particularly if they were alone. At the museum, there was clearly more headroom for expansive and un-impeded behaviour. Many of the permanent museum exhibitions are designed for interaction and exploration, and the wide variety of strange projects taking part in the Maker Faire clearly made people less concerned about how their behaviour would be perceived by others, as this behaviour was expected in this context. Nevertheless, there were examples at both locations of people showing an interest in the installation but being too shy to try it for themselves. But by having the opportunity to watch others interact with it and build an understanding of how it worked, the shyness was sometimes overcome by curiosity, i.e., the impetus to engage became stronger than the impedance.

Our initial approach to grasp differences between the two locations and experiences people had while interacting with the installation in the library or at the Makers Faire, was to use grounded theory and coding. The main method was passive observation with coding. Three coders were used, two in the library, and a third person in the museum. This was done in order to minimize subjectivity. Furthermore, the coding schemas and outputs were compared and discussed among the three coders. Only data where there was agreement between all three was taken into account. Some measurements, such as taking time, were relatively straightforward. The main difficulty there was in keeping track of people getting in and leaving interaction space during one interactive session. The other measurements were more complex, such as recording people's facial expressions and body language.

In terms of time spent by participants interacting with the installation in the two locations, we found out that the time spent at the museum was significantly higher. At the library, no one spent more than three minutes with the installation, 41% spent less than one minute and 72% of the observed spent two minutes or less. At the museum, the time spent with the installation is spread much more evenly across the intervals noted: 59% spent two minutes or more interacting, and some people seen outside the time frame of observations were exceeding the intervals noted significantly.

Looking at the distribution of facial expressions observed in the two different contexts (Fig. 12 and Fig. 13), expressions of a positive nature are the predominant ones in both settings, but at the museum as many as 86% were smiling and even though 5% were noted as indifferent, 95% of the observed were deemed to have a positive experience.

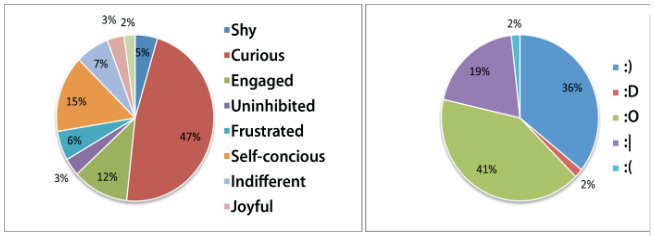


Figure 12. Body language and facial expression distributions at the library.

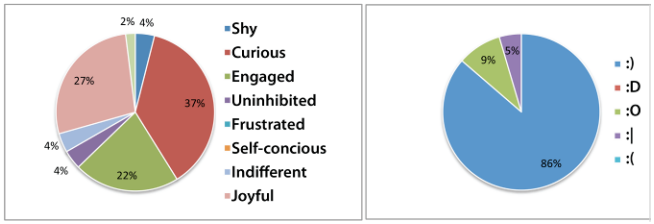


Figure 13. Body language and facial expression distribution at the museum.

Comparing the observations of body language between the contexts, a high degree of curiosity is observed in both settings, with 47% recorded as displaying a body language suggesting curiosity in the library setting, while 37% were recorded at the museum. The most striking difference between the library and the museum contexts was the high percentage of joyfulness (27%) and the low percentage of shyness (4%) of the museum setting, contrasting the low degree of joyfulness (3%) and high degree of self-consciousness (15%) and shyness (5%) (combined 20%) seen at the library. The reason for combining self-consciousness and shyness is that they are very similar traits. Seen in retrospect, separating these terms into two coding categories might have been unnecessary, considering their similarities and the fallibility of observation.

If we look at the distribution of the differences observed in the body language, it seems that the library context was perceived as a less comfortable one. The given percentages may be directly related to impetus (curiosity, joyfulness, engagement) and impedance (shyness, self-consciousness, inhibitions), as these various attributes are just diverse

aspects of impetus or impedance. What we can then see at once is that nearly a third (33%) of participants in interactions at the library experienced some form of impedance, vs. only 8% in the museum.

Whole 95% of people who participated in the museum had expressions of satisfaction, vs. only 79% in the library. On the other hand, the percentage of those who found themselves fascinated was 41% in the library, vs. only 9% in the museum. This might suggest that fascination was expected to happen in the museum, while it was unexpected in the library.

VI. Discussion of the General Findings

The previous paragraph described our best effort at the time to evaluate suitability of spaces for our installation and measure differences in terms of the time spent on the installation at different locations. Here, we take a look at how considering ecologies of interactive spaces (with space, people, technology, activity and values as main components) through the use of concepts of immediacy, impetus, impedance, and the fluidity of sharing, may give a richer perspective on design and evaluation of installations for enjoyment in public spaces.

A. Immediacy

Entering the library, the installation, unless in use, was not immediately noticeable. Thus, the space itself did not provide for immediacy. As noted earlier, the visitors were used to information screens that were used for the installation, and physical changes to the space were minimal.

Display blindness is a term used to describe the phenomenon where people can selectively ignore screens. *Interaction blindness* refers to the fact that it is difficult for people to understand whether a given display is interactive. Houben and Weichel [33] have described how display blindness and interaction blindness can be overcome by use of curiosity objects, e.g., objects that are designed to draw attention by sparking interest and curiosity. The term we use, impetus, is closely related to curiosity objects, being defined as all that (by design) nudges curiosity, interest, and activity.

When the installation was in use, people entering the library would have immediate understanding that something different is happening in the open area. The music and the interactive movements are, normally, not part of the library experience.

The situation was different at the museum during the Makers Faire. The installation was the only activity happening in the room. It was immediately understandable that there was an activity available in the room by observing the triangle on the floor, the lit-up screen and the sitting blocks (see Fig. 6). Thus, even if the installation was not in use at the moment of the entry to the room, the understanding that some activity is available in the room was immediate. An overview of the situation in the room was available at a glance, whether the installation was in use or not.

The two different spaces that were used for the installation within the library illustrate the importance of considering how immediacy, in relation to understanding the space, influences interaction with installation.

Immediacy can also be used to address the interface and its properties: is it easy to understand what one should do to engage with the system? Are activities in the space properly understood at the glance?

While it is not clear how to apply immediacy to values, one can pose the question when designing for a specific location: who are the people frequenting the location and what are their values? Is this information available at a glance at this location, or, are these aspects something that has to be found out gradually?

B. Impetus

The sound, graphical design and body movements are considered to be the most important, designed, ways of providing impetus for this installation. Perhaps in contrast to the usual absence of music in the library, people who were sitting in the open area, for example, in the café, soon learned that whenever the music started, there were people interacting with the installation. This allowed them to look up whenever the installation was in use, thereby slowly building an understanding of how it worked. This also allowed them to build both curiosity and courage to try the installation for themselves. We saw several examples of people coming up to investigate after having observed others interacting with it for a while. There were also examples of single persons and groups of people who were hanging around in the background, queuing when others were interacting with the installation. As soon as the people using the installation left, they would walk up and give it a try. This worked like a honeypot effect, a positive feedback loop, where use attracted attention and instigated more use. So use was also an impetus, a call for engagement. However, the installation was unable to keep people's interest for more than a minute or two, which meant that there would have to be a constant stream of people to keep the installation in continuous use. When the installation was allowed to go into standby mode, people quickly returned their attention to whatever they were otherwise doing.

Further impetus was provided for by-passers by starting the system whenever someone came into the detection range for Kinect devices.

In the museum, further nudging was provided by clearly marking the interactive space on the floor, so it was easy to understand that there was enough space for more than one person, and this feature has enabled more group interactions than we observed in the library, also with total strangers, see Fig. 11. It was also clear that the norms and values in two places were different, the level of impetus that people needed, was lower in the museum than in the library.

Impetus, thus, can be a part of design considerations when developing the interface, reflecting over the physical location of the installations, and as part of the activities the installation provides. If successful, the use of installation increases the positive effects.

C. Impedance

The two locations at the library were exposed and crowded, particularly the first one. This meant that anyone interacting with the installation would draw attention from not only the immediate surroundings, but also from galleries on the floors above. The sounds naturally draw attention from the surroundings. Thus, for many people this attention from the surroundings is not desirable and prevented them from engaging with the installation. The people who interacted with the system in the library on the third day, in a bit more protected area, often took the elevator shaft as some sort of extra protection from onlookers, see Fig. 8.

Other impeding factors in the library were related to the activities, the lack of time, sense of the work environment, as well as the norms and values related to the space.

The space we were assigned at the museum during the Maker Faire was partially confined, making it close to impossible for others to observe the installation, or the people interacting with it, from afar. This seemed to give participants a sense of privacy and allowed them to let themselves get more carried away than at the library. Also, having sitting blocks for onlookers to sit on was very beneficial, Fig. 5. It allowed the ones who did not want to try the installation to sit down and relax, but still be able to communicate and take part in the experience with their friends who were interacting with the installation. Thus, impedance was minimized by providing sense of privacy, safety, some level of comfort and ability to participate, even when sitting on the sidelines.

D. Fluidity of sharing

The system used for our installation was designed with fluidity of sharing in mind. Kinect naturally enables multiple users to interact with the system, but our application used abstract graphical interfaces, allowing, ideally, everyone to enjoy interacting and sharing. In addition, we observed that the fluidity of sharing was much higher in the museum. There, we could observe strangers interacting (see Fig. 11) along the side of those who were friends or family, while the group interactions in the library involved mostly friends. This suggests that fluidity also depends on the norms and values of the space.

VII. Conclusion

We have defined the ecology of interactive spaces as a function of the physical properties of the space, people, activities, technologies and values and norms associated with the location for the installation. Looking at the ecology of interactive spaces is both timely, as public space interactions are becoming ubiquitous, and desirable since concepts for evaluating experiences are still few and divided by disciplines, e.g., social sciences, HCI. Combining the technical, social, architectural and human aspects of the space for which interaction is designed, significantly increases chances to succeed in creating enjoyable interactive installations in the public room. In order to facilitate reflection and design of ecologies of interactive spaces, we have introduced concepts of immediacy, impetus, impedance and fluidity of sharing as aid in "getting the big picture" first. We believe that if we had this framework at the start of our own design process, the outcome of that process would have been different, and better. It would aid understanding and study of relationships between the space, people, technology, interaction with it and values.

Our hope is that the concept will grow and get to be better defined through other examples and studies, yielding a set of principles and guidelines not only for design, but also for the evaluation of interactive installations in public spaces.

References

[1] R. Rosseland, S. Berge, and A. L. Culén, "Publicly Displayed Interactive Installations: Where Do They Work Best?" Proceedings of the Seventh International Conference on Advances in Computer-Human Interactions, 2014, pp. 1–8.

[2] "Design I/O - Funky Forest Moomah." [Online]. Available from: http://design-io.com/projects/Moomah/. Accessed on Dec. 5, 2014.

[3] M. Baas, C. K. W. De Dreu, and B. A. Nijstad, "A meta-analysis of 25 years of mood-creativity research: hedonic tone, activation, or regulatory focus?" Psychol. Bull., vol. 134, no. 6, pp. 779–806, Nov. 2008.

[4] S. Doorley and S. Witthoft, Make Space: How to Set the Stage for Creative Collaboration, 1st edition. Wiley, 2012.

[5] A. P. McGinn, K. R. Evenson, A. H. Herring, S. L. Huston, and D. A. Rodriguez, "Exploring Associations between Physical Activity and Perceived and Objective Measures of the Built Environment," J. of Urban Health, vol. 84, no. 2, pp. 162–184, Mar. 2007.

[6] "Do Google's playful perks spark creativity?" SmartPlanet. [Online]. Available from: http://www.smartplanet.com /blog/bulletin/do-googles-playful-perks-spark-creativity/. Accessed on Dec. 5, 2014.

[7] B. A. Nardi and V. L. O'Day, Information ecologies: using technology with heart. Cambridge, Mass., MIT Press, 1999.

[8] V. Kaptelinin and L. J. Bannon, "Interaction Design Beyond the Product: Creating Technology-Enhanced Activity Spaces," Human–Computer Interaction, vol. 27, no. 3, pp. 277–309, 2012.

[9] E. Hornecker and J. Buur, "Getting a grip on tangible interaction: a framework on physical space and social interaction," Proceedings of the SIGCHI conference on Human Factors in computing systems, 2006, pp. 437–446.

[10] E. Hornecker, P. Marshall, and Y. Rogers, "From Entry to Access: How Shareability Comes About," Proceedings of the Conference on Designing Pleasurable Products and Interfaces, New York, NY, USA, 2007, pp. 328–342.

[11] K. Battarbee, "Defining Co-experience," Proceedings of the International Conference on Designing Pleasurable Products and Interfaces, New York, NY, USA, 2003, pp. 109–113.

[12] J. Forlizzi and K. Battarbee, "Understanding experience in interactive systems," Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques, 2004, pp. 261–268.

[13] M. T. Koppel, G. Bailly, J. Müller, and R. Walter, "Chained displays: configurations of public displays can be used to influence actor-, audience-, and passer-by behaviour," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA, 2012, pp. 317–326.

[14] C. Heath and D. von Lehn, "Configuring 'Interactivity' Enhancing Engagement in Science Centres and Museums," Social Studies of Science, vol. 38, no. 1, pp. 63–91, Feb. 2008.

[15] M. Blythe and M. Hassenzahl, "The Semantics of Fun: Differentiating Enjoyable Experiences," Funology, M. A. Blythe, K. Overbeeke, A. F. Monk, and P. C. Wright, Eds. Springer Netherlands, 2005, pp. 91–100.

[16] M. Csikszentmihalyi, Flow: the psychology of optimal experience. New York, N.Y.: Harper Perennial, 1991.

[17] B. Sutton-Smith, The ambiguity of play. Cambridge, Mass., Harvard University Press, 1997.

[18] K. Salen and E. Zimmerman, Rules of play: game design fundamentals. Cambridge, Mass., MIT Press, 2003.

[19] R. Caillois, Man, Play, and Games. University of Illinois Press, 2001.

[20] B. Costello and E. Edmonds, "A Tool for Characterizing the Experience of Play," Proceedings of the Sixth Australasian Conference on Interactive Entertainment, New York, NY, USA, 2009, pp. 2:1–2:10.

[21] B. Costello and E. Edmonds, "A study in play, pleasure and interaction design," Proceedings of the conference on Designing pleasurable products and interfaces, New York, NY, USA, 2007, pp. 76–91.

[22] J. N. Lieberman, Playfulness: its relationship to imagination and creativity. New York: Academic Press, 1977.

[23] J. Ferrara, Playful Design, 1st edition. Rosenfeld Media, 2012.

[24] H. Gardner, Frames of Mind: The Theory of Multiple Intelligences. New York. Basic Books, 1983.

[25] A. Karabeg, M. N. Akkok, and K. Kristensen, "Towards a language for talking about information visualization aimed at presentation on the Web," Proceedings of the Eighth International Conference on Information Visualisation, IV 2004, 2004, pp. 930 – 937.

[26] A. L. Culén, "Visual Immediacy for Sense-Making in HCI," Proceedings of the international Conference on Interfaces and Human Computer Interaction, 2014, pp. 265–270.

[27] F. Garzotto and F. Rizzo, "Interaction Paradigms in Technology-enhanced Social Spaces: A Case Study in Museums," Proceedings of the Conference on Designing Pleasurable Products and Interfaces, New York, NY, USA, 2007, pp. 343–356.

[28] R. Castro, "Var det alt? En studie av brukeropplevelser i TV-studio på INSPIRIA science center," M.S. Thesis, University of Oslo, 2014.

[29] S. C. Bolton and M. Houlihan, "Are we having fun yet? A consideration of workplace fun and engagement," Employee Relations, vol. 31, no. 6, pp. 556–568, Oct. 2009.

[30] K. Krippendorff, The Semantic Turn: A New Foundation for Design. Boca Raton: CRC Press, 2005.

[31] "The Royal London Hospital Play Space | Room to Bloom." [Online]. Available from: http://www.room-to-bloom.com /blog/the-royal-london-hospital-play-space/. Accessed on Dec. 5, 2014.

[32] A. Maimone and H. Fuchs, "Reducing interference between multiple structured light depth sensors using motion," Proceedings of IEEE Virtual Reality Conference, VRW 2012, 2012, pp. 51–54.

[33] S. Houben and C. Weichel, "Overcoming interaction blindness through curiosity objects," Extended Abstracts on Human Factors in Computing Systems, CHI '13, New York, NY, USA, 2013, pp. 1539–1544.

[34] "Go With The Flow." [Online]. Available from: http://www.wired.com/wired/archive/4.09/czik_pr.html. Accessed on Dec. 5, 2014.

[35] B. M. Costello and E. A. Edmonds, "Directed and emergent play," Proceedings of the Seventh ACM conference on Creativity and cognition, New York, NY, USA, 2009, pp. 107–116.

[36] S. Harrison, D. Tatar, and P. Sengers, "The three paradigms of HCI," in Alt. Chi. Session at the SIGCHI Conference on Human Factors in Computing Systems, San Jose, California, USA, 2007, pp. 1–18.

[37] A. L. Culén and A. Gasparini, "Situated Techno-Cools: factors that contribute to making technology cool and the study case of iPad in education," PsycNology Journal, vol. 10, no. 2, pp. 117–139, 2012.

Legacy Network Infrastructure Management Model for Green Cloud Validated Through Simulations

Sergio Roberto Villarreal, María Elena Villarreal, Carlos Becker Westphall, and Carla Merkle Westphall
Network and Management Laboratory – Post-Graduate Program in Computer Science
Federal University of Santa Catarina
Florianopolis, SC, Brazil
sergio@lrg.ufsc.br, maria@lrg.ufsc.br, westphal@lrg.ufsc.br, carla@lrg.ufsc.br

Abstract — **The concepts proposed by Green IT have changed the priorities in the design of information systems and infrastructure, adding to traditional performance and cost requirements, the need for efficiency in energy consumption. The approach of Green Cloud Computing builds on the concepts of Green IT and Cloud in order to provide a flexible and efficient computing environment, but their strategies have not given much attention to the energy cost of the network equipment. While Green Networking has proposed principles and techniques that are being standardized and implemented in new networking equipment, there is a large amount of legacy equipment without these features in data centers. In this paper, the basic principles pointed out in related work for power management in legacy network equipment are presented, and a model for its use to optimize green cloud approach is proposed. It is also presented NetPowerCloudSim, an extension to the open-source framework CloudSim, which was developed to validate the aforementioned model and adds to the simulator the capability of representing and managing network equipment according to the state changes of servers. Experiments performed to validate the model showed that it is possible to significantly increase the data center efficiency through its application. The major contributions of this paper are the proposed network infrastructure management model and the simulator extension.**

Keywords - Green IT; Cloud Computing; Network Management; Data center; CloudSim.

I. INTRODUCTION

This paper extends [1], which proposes a data center's network equipment management model to optimize the green cloud approach, presenting an extension to the CloudSim simulator and the experiments performed to validate the aforementioned model.

Traditionally, computer systems have been developed focusing on performance and cost, without much concern for their energy efficiency. However, with the advent of mobile devices, this feature has become a priority because of the need to increase the autonomy of the batteries.

Recently, the large concentration of equipment in data centers brought to light the costs of inefficient energy management in IT infrastructure, both in economic and environmental terms, which led to the adaptation and application of technologies and concepts developed for mobile computing in all IT equipment.

The term Green IT was coined to refer to this concern about the sustainability of IT and includes efforts to reduce its environmental impact during manufacturing, use and final disposal.

Cloud computing appears as an alternative to improve the efficiency of business processes, since from the point of view of the user, it decreases energy costs through the resources sharing and efficient and flexible sizing of the systems. Nevertheless, from the standpoint of the service provider, the actual cloud approach needs to be seen from the perspective of Green IT, in order to reduce the data center energy consumption without affecting the system's performance. This approach is known as Green Cloud Computing [2].

Considering only IT equipment, the main cause of inefficiency in the data center is the low average utilization rate of the resources, usually less than 50%, mainly caused by the variability of the workload, which obliges to build the infrastructure to handle work peaks that rarely happen, but that would decrease the quality of service if the application was running on a server fully occupied [3].

The strategy used to deal with this situation is the workload consolidation that consists of allocating the entire workload in the minimum possible amount of physical resources to keep them with the highest possible occupancy, and put the unused physical resources in a state of low energy consumption.

The challenge is how to handle unanticipated load peaks and the cost of activation of inactive resources. Virtualization, widely used in the Cloud approach, and the ability to migrate virtual machines have helped to implement this strategy with greater efficiency.

To validate green cloud management algorithms and strategies, simulators are used, since performing tests in real environments is not feasible due to the cost, the physical rigidity of the structure and the difficulty to reproduce experiments under controlled conditions.

Calheiros et al. [4] developed CloudSim, an open-source framework for modeling and simulating cloud computing environments, which allows performing simulations of large scale data center operation in a conventional computer. With

this simulator, it is possible to conduct experiments to validate workload consolidation algorithms, measure power consumption and calculate violations to the hired service levels. However, its use demands effort to interpret the code and extend it.

Strategies to improve efficiency in data centers have been based mainly on the servers, cooling systems and power supply systems, while the interconnection network, which represents an important proportion of energy consumption, has not received much attention, and the proposed algorithms for load consolidation of servers, usually disregard the consolidation of network traffic [5][6].

According to Bianzino et al. [7], traditionally the networking system design has followed two principles diametrically opposed to the aims of Green Networking: oversizing to support demand peaks and redundancy for the single purpose of assuming the task when other equipment fails.

The concepts of Green IT, albeit late, have also achieved design and configuration of network equipment, leading to Green Networking, which primary objective is to introduce the concept of energy-aware design in networks without compromising performance or reliability, and has to deal with a central problem: the energy consumption of traditional network equipment is virtually independent of the traffic workload [8].

The Green Networking has as main strategies proportional computing that applies to adjust both the equipment processing speed and the links speed to the workload, and the traffic consolidation, which is implemented considering traffic patterns and turning off not needed components.

While the techniques of Green Networking begin to be standardized and implemented in the new network equipment, a large amount of legacy equipment forms the infrastructure of current data centers. In works to be presented in the next section, it is shown that it is possible to manage properly these devices to make the network consumption roughly proportional to the workload.

Taking into account that the more efficient becomes the management of virtual machines and physical servers, the greater becomes the network participation in the total consumption of the data center, the need to include network equipment in green cloud model is reinforced.

Thereby, there is the need and the possibility to add, to the Green Cloud management systems, means of interaction with the data center network management system, to synchronize the workload consolidation and servers shutdown, with the needs of the network traffic consolidation.

In this article, the principles suggested in recent papers by several authors for power management in legacy network equipment are presented, and their application to optimize green cloud approach is proposed. An extended version of the CloudSim called NetPowerCloudSim and the results of the experiments performed to validate the model are also presented.

The remainder of this paper is organized as follows: Section II describes related work on which is based our proposal that is presented in Section III, along with an analytic case study to show the model's application possible results. Section IV presents NetPowerCloudSim, the experiments performed with this extended simulator to validate the model, and the results obtained. Finally, in Section V, concluding remarks and proposals for future work are stated.

II. Related Work

Mahadevan et al. [9] present the results of an extensive research conducted to determine the consumption of a wide variety of network equipment in different conditions. The study was performed by measuring the consumption of equipment in production networks, which made it possible to characterize the energy expenditure depending on the configuration and use of the equipment, and determine a mathematical expression that allows calculating it with an accuracy of 2%. This expression determines that total consumption has a fixed component, which is the consumption with all ports off, and a variable component which depends on the number of active ports and the speed of each port.

Research has determined that the power consumed by the equipment is relatively independent of the traffic workload and the size of packets transmitted, and dependent on the amount of active ports and their speed. The energy saved is greater when the port speed is reduced from 1 Gbps to 100 Mbps, than from 100 Mbps to 10 Mbps.

This research also presents a table with the average time needed to achieve the operational state after the boot of each equipment category, and also demonstrates that the behavior of the current equipment is not proportional, as expected according to the proposals of the Green Networking, and therefore the application of traffic consolidation techniques have the potential to produce significant energy savings.

Mahadevan et al. [10], continuing the work presented in the preceding paragraphs, put the idea that the switches consumption should ideally be proportional to the traffic load, but as in legacy devices the reality is quite different, they propose techniques to make the network consumption closer to the proportional behavior by the application of configurations available in all devices.

The results are illustrated in Figure 1, which shows the ideal behavior identified as "Energy Proportional" which corresponds to a network with fully "Energy Aware" equipment, the actual curve of the most of the today's networks where the consumption is virtually independent of load, labeled "Current", and finally the consumption curve obtained by applying the techniques they proposed, labeled "Mahadevan's techniques".

The recommended configurations are: slow down the ports with low use, turn off unused ports, turn off line cards that have all their ports off and turn off unused switches. The authors, through field measurements, have shown that it

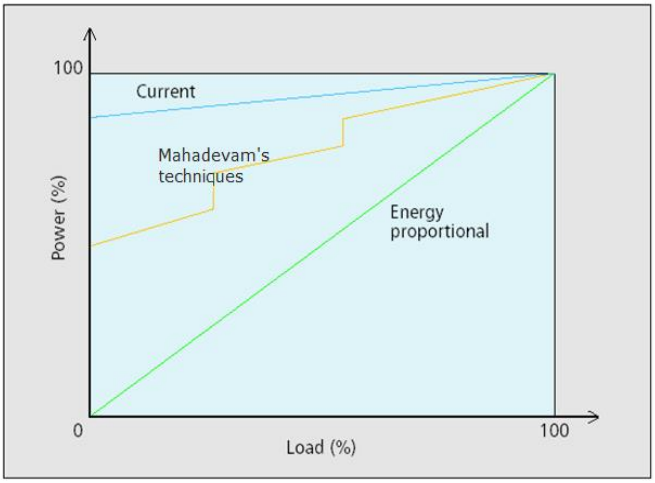


Figure 1. Consumption in computer networks as a function of the workload [10].

is possible to obtain savings of 35% in the consumption of a data center network with the application of these settings. Also, with the use of simulations, they have demonstrated that in ideal conditions savings of 74% are possible combining servers load consolidation and network traffic consolidation.

Werner [11] proposes a solution for the integrated control of servers and support systems for green cloud called OTM (Organization Theory Model). This approach, based on the Theory of Organization, defines a model of allocation and distribution of virtual machines that were validated through simulations and showed to get up to 40% energy saving compared to traditional cloud model.

The proposed model determines when to turn off, resize or migrate virtual machines, and when to turn on or off physical machines based on the workload and the SLA (Service Level Agreement) requirements. The solution also envisages the shutdown of support systems. Figure 2 shows the architecture of the management system proposed, which is based on norms, roles, rules and beliefs.

Calheiros et al. [4], from the University of Melbourne, present the CloudSim simulator, an open-source framework which supports large scale cloud environment modeling and simulation in a conventional computer with low consumption of computational resources. This tool was designed specifically for modeling cloud computing infrastructures, and offers support to virtualized environments simulation and to modeling data centers with large amounts of servers. This version of the simulator has a class called NetworkTopology, which provides information about entities communication latency, but does not allow representing network equipment and energy consumption.

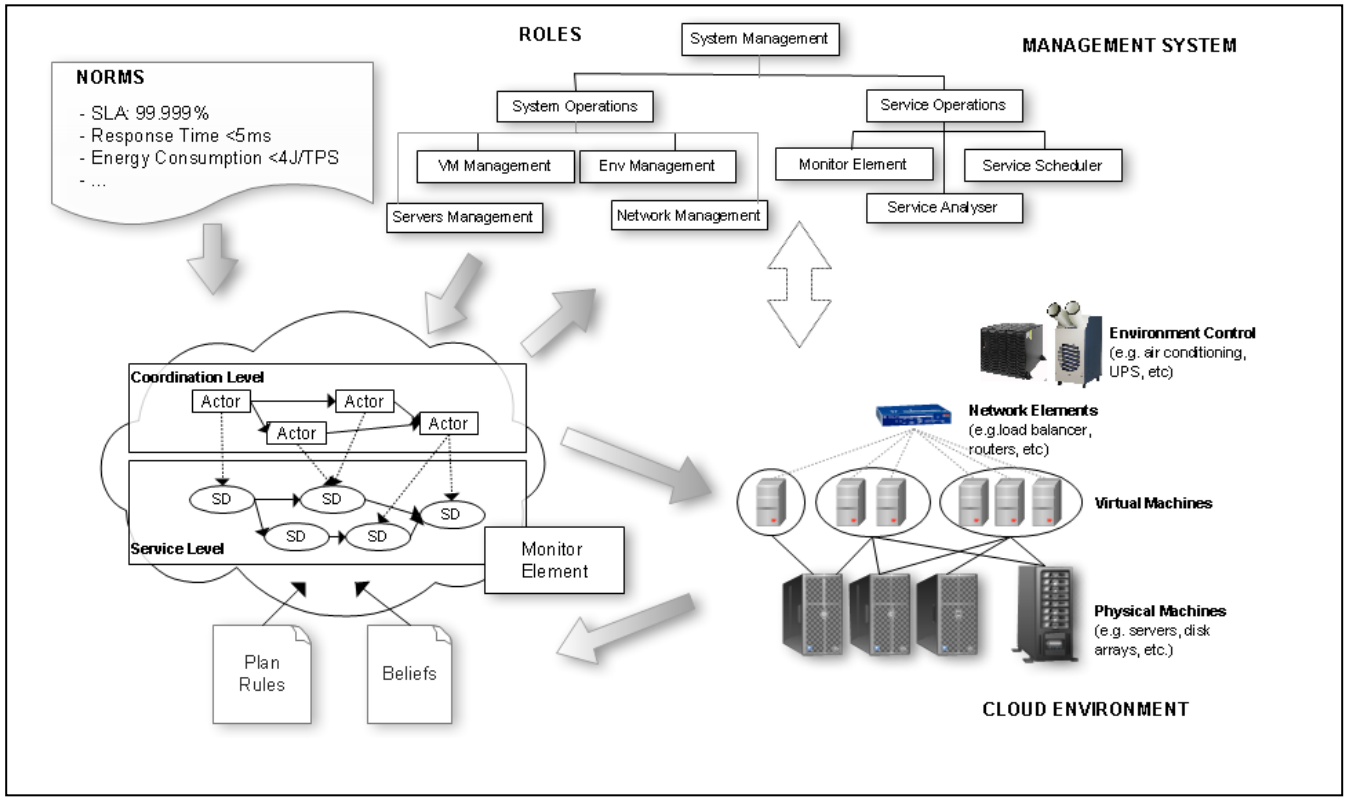


Figure 2. Green Cloud management system based on OTM [11].

Freitas [12] made extensions to the CloudSim simulator, creating the needed classes to support and validate the OTM, which allowed calculating energy savings and SLA violations in various scenarios. Neither the model nor the extensions consider the network equipment energy consumption.

Garg and Buyya [13] present NetworkCloudSim, an extension to CloudSim, which incorporates resources for modeling applications and data center network behaviors. This simulator has the necessary classes to represent network equipment and traffic, however, it does not allow representing and calculating data center equipment energy consumption.

Beloglazov [14] presented a new version of the simulator, the CloudSim 2.0, which allows representing the data center components energy consumption, capacity that was not contemplated by the framework core, and also incorporates applications with dynamic workloads. This version does not support network equipment and its energy consumption representation.

Based on the findings of the works described above, in the next section, a proposal to include the management of legacy and current network devices in OTM is presented. The rules and equations required to include this extension in CloudSim simulations are also presented and validated through a case study.

III. PROPOSAL FOR DATA CENTER NETWORK MANAGEMENT IN GREEN CLOUD APPROACH

The proposal considers the network topology of a typical data center shown in Figure 3, where the switches are arranged in a hierarchy of three layers: core layer, aggregation layer and access or edge layer. In this configuration, there is redundancy in the connections between layers so that the failure of a device does not affect the connectivity.

In traditional facilities, the implementation and management of this redundancy is done by the Spanning Tree Protocol and in most recent configurations by the MC-LAG (Multichassis Links Aggregation Group), which allows using redundant links simultaneously expanding its capacity, as described in [15].

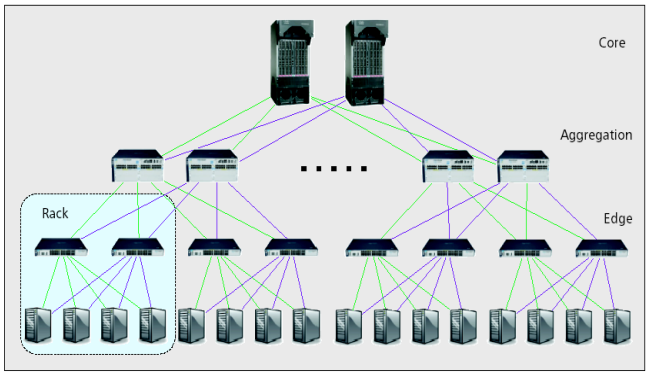


Figure 3. Typical network topology of a data center [9].

The racks are the basic unit of this configuration and each rack accommodates a certain amount of servers and two access layer switches. The servers have two NICs (Network Interface Card) each one connected to a different access switch.

A. Extensions To The Organization Theory Model

To include the management of legacy network equipment in the model proposed by Werner et al. [16], such that the network consumption becomes relatively proportional to the traffic workload and the energy savings contribute to the overall efficiency of the system, it is proposed to add the following elements to its architecture:

1) Management Roles

Add to the "System Operations" components the "Network Equipment Management" role, which acts as an interface between the model and the network equipment being responsible for actions taken on these devices such as: enabling and disabling ports or equipment or change MC-LAG protocol settings.

The "Monitoring Management" role, responsible for collecting structure information and its understanding, should be augmented with elements for interaction with the network management system to provide data, from which decisions can be made about the port speed configuration, or turning on or off components and ports. These decisions will be guided by the rules and beliefs.

2) Planning Rules

These rules are used when decisions must be taken, and therefore, rules to configure the network equipment in accordance with the activation, deactivation and utilization of physical machines should be added.

To implement the settings pointed out in [9], already presented, the following rules are proposed:

- If a PM (Physical Machine) is switched off, the corresponding ports of access layer switches must be turned off.
- If the occupation of a PM is smaller than a preset value, network interfaces and corresponding access switches ports must be slowed down.
- If the aggregate bandwidth of the downlink ports of an access layer switch is smaller than a preset value, their uplink ports must have their speed reduced.
- If an access layer switch has all its ports off, it must be turned off.
- If an access layer switch is turned off, the corresponding ports of the aggregation layer switch must be turned off.
- If the aggregate bandwidth of the downlink ports of an aggregation layer switch is smaller than a preset value, their uplink ports must have their speed reduced.
- If an aggregation layer switch has all its ports off, it must be turned off.
- If an aggregation layer switch is turned off, the corresponding port of the core layer switch must be turned off.

- If a module of a core layer switch has all its ports off, it must be turned off.
- If a core layer switch has all its ports off, it must be turned off.
- All reversed rules must also be included.

The application of these rules does not affect the reliability of the network, since port and devices are only turned off when servers are turned off. The system performance will only be affected if the network equipment activation cost is bigger than the server activation cost.

For more efficiency in traffic consolidation, the model should consider the racks in virtual machines allocation and migration strategies, and rules that consolidate active physical machines in as fewer racks as possible are necessary.

3) Beliefs

They are a set of empirical knowledge used to improve decisions, and are linked to the used resources characteristics and to the type of services implemented in each specific case.

For each of the rules listed in the previous paragraph, a belief related to energy consumption should be stated. If we consider Christensen et al. [17], examples include:

- Disconnecting a port of an access layer switch generates a saving of 500 mWh.
- Decreasing the speed of a port from 10 Gbps to 1 Gbps generates a saving of 4.5 Wh.

It will also be necessary to include beliefs about the time required for a deactivated port or device to become operational after the boot. These beliefs will be used to make decisions that must consider performance requirements.

B. Simulation Model

The typical data center network topology, rules and beliefs proposed form the basis for building a simulation model to validate different strategies and rules in specific settings and with different workloads.

For the simulator implementation, it was considered that each rack accommodates forty 1U servers and two access layer switches. Each of these switches has 48 Gigabit Ethernet ports and two 10 Gigabit Ethernet uplink ports. Each server has two Gigabit Ethernet NICs (Network Interface Card) each one connected to a different access switch.

It was also considered that if there is only one rack, aggregation layer switches are not required, and up to 12 racks can be attended by 2 aggregation layer switches with twenty four 10 Gigabit Ethernet and two 10 Gigabit Ethernet or 40 Gigabit Ethernet uplinks, with no need for core switches.

Finally, it was assumed that, with more than 12 racks two core switches with a 24 ports module for every 144 racks will be required. The module's port speed may be 10 Gigabit Ethernet or 40 Gigabit Ethernet, according to the aggregation switches uplinks.

In the next subsections the central aspects of the simulation model are presented.

1) Network Topology Definition

The simulator must create the network topology based on the amount of physical servers using the following rules:

- If the number of servers is smaller than 40, the topology will have only two access layer switches interconnected by their uplink ports. Turn off unused ports.
- If the number of servers is greater than 40 and smaller than 480 (12 Racks), put two access layer switches for every 40 servers or fraction and two aggregation layer switches interconnected by their uplink ports. Turn off unused ports of both layers switches.
- If the number of servers is greater than 480, apply the previous rule for each group of 480 servers or fraction, add two core layer switches and put on each switch a 24 ports module for each 5,760 servers (144 racks) or fraction. Turn off unused port.

2) Network Energy Consumption Calculation

The total consumption of the network is given by the sum of all its switches consumption and, based on the findings of Mahadevan et al. [9], the equation to calculate switches and modules consumption is:

$$\text{Power (W)} = \text{BP} + \text{no. P 40Giga} \times 10 + \text{no. P 10Giga} \times 5 + \text{no. P Giga} \times 0.5 + \text{no. P Fast} \times 0.3 \quad (1)$$

In this expression, the power in Watts is calculated by summing the base power (BP), which is a fixed value specific to each device, and the consumption of every active port at each speed, which is a variable component. The consumption of each type of port is specific to each device, but the proposed values are the average values according to the works already cited.

The simulator must permit to set each kind of port consumption and the BP in order to represent different scenarios and to calibrate the model.

In (1), if the switch is modular, the base power of the chassis must be added.

At the end of each simulation frame, the simulator must update the calculation of the network total consumption by summing each switch consumption during the frame.

3) Interconnection calculation

Since the network topology is a hierarchy, it is possible to establish a mathematical relationship in the equipment interconnection if these are identified by numbers. Thus, it is not necessary to include information about these interconnections in the state vector.

When the simulator needs to determine the switch port number that corresponds to a specific server network interface, or to a specific uplink port of a switch, it is possible to calculate it using a mathematical expression applied to the server or switch identifier.

4) Network Management

During the simulation, when servers are connected or disconnected, the simulator must apply the network management rules by turning on or off the corresponding ports or configuring its speed.

The sequence of the application of the rules according to the state changes of servers is represented by the activity diagram in Figure 4. This diagram considers, besides the events of turning servers on and off, events based on the utilization rate of the server.

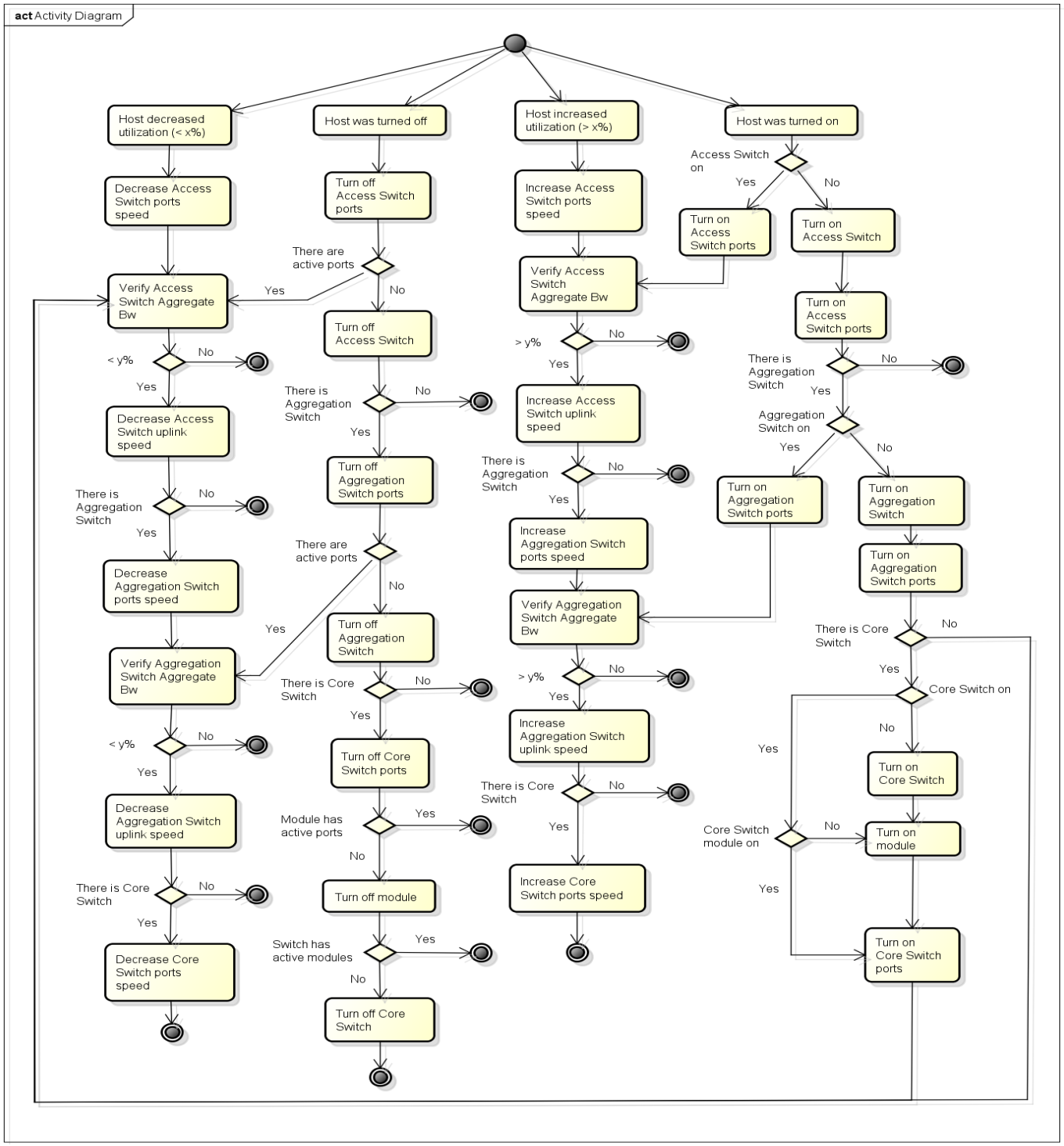


Figure 4. Activity Diagram of the application of the model rules.

C. Case Study

To validate the model and the potential of the proposal, it was applied to a hypothetical case of a cloud with 200 physical servers, creating the topology, calculating its initial consumption without network equipment management and illustrating two possible situations in the operation of the system. It was considered for this scenario that the base power is 60 W for access layer switches and 140 W for aggregation layer switches.

Applying the rule to calculate the topology, it is determined that it comprises 5 racks housing a cluster of 40 servers each and, therefore, there will be ten access layer switches with forty Gigabit Ethernet ports and two 10 Gigabit Ethernet empowered ports, and two aggregation layer switches with ten 10 Gigabit Ethernet connected ports for access layer switches and two 40 Gigabit Ethernet ports for uplink interconnection between them.

1) Scenario 1: All network equipment with all its ports connected

Access layer switches = 10 x (60 + 2x5 + 48x0.5)
= 940 W
Aggregation layer switches = 2 x (140 + 2x10 + 24x5)
= 560 W
Total network consumption = 1,500 W

2) Scenario 2: Initial configuration with unused ports off

Access layer switches = 10 x (60 + 2x5 + 40x0.5)
= 900 W
Aggregation layer switches = 2 x (140 + 2x10 + 10x5)
= 420 W
Total network consumption = 1,320 W

In this scenario, it is observed that only by the proper initial configuration of the network it is possible to get a power save of approximately 12%.

3) Scenario 3: 90 active servers, workload consolidated in the first three racks and network configuration rules applied.

In this situation, according to the rules, there are 4 access layer switches working in initial conditions (2), two access layer switches working with twelve Gigabit Ethernet ports, 10 for servers and 2 uplink ports with its speed reduced (3), and 2 aggregation layer switches with four 10 Gigabit Ethernet and two Gigabit Ethernet downlinks ports and two 40 Gigabit Ethernet uplinks (4), and the network consumption will be:

Access layer switches 1 = 4 x (60 + 2x5 + 40x0.5)
= 360 W (2)
Access layer switches 2 = 2 x (60 + 12x0.5)
= 132 W (3)
Aggregation switches = 2 x (140 +2x10 + 4x5 + 2x0.5)
= 362 W (4)
Total network consumption = 854 W

In this scenario, there is a power saving of approximately 43% in network consumption.

IV. NETPOWERCLOUDSIM

To validate the network management model proposed in the previous section, extensions to the CloudSim were developed and the extended simulator was called NetPowerCloudSim.

A. Extensions development

To represent the network and manage it according to the rules of the model, the PowerSwitch, NetTopology, NetworkManager and NetPowerDatacenter classes were developed as presented in the simplified class diagram in Figure 5.

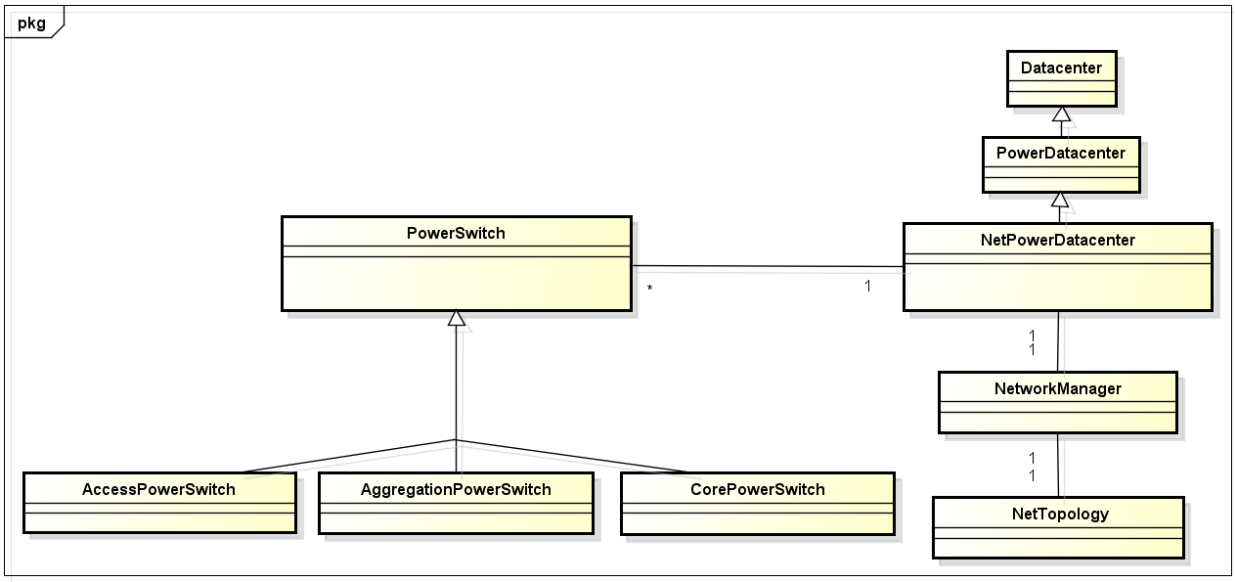


Figure 5. NetPowerCloudSim simplified class diagram.

The PowerSwitch class represents network equipment and is extended by other classes that represent specific layer switches (access, aggregation and core). The switches have attributes such as status (on/off), current consumption, quantity of ports and each port's speed; and methods that allow to turn them on and off, to set the speed of a specific port and to calculate their consumption at a given time, as well as the energy consumed during a simulation frame, which is done through linear interpolation.

The NetTopology class represents the network topology and is responsible for calculating the quantity of each kind of switch and their interconnection. Before the simulation start, based on the data center physical machines amount, this class calculates the number of racks needed to accommodate them. The quantities of access, aggregation and core switches are then calculated from the amount of racks.

The NetworkManager class contains the attributes and the logic required to turn on and off network equipment and ports, and to set ports speed when the state of servers changes, based on the management model rules and beliefs. Before the simulation start, helped by the NetTopology class, it determines which ports are not connected to any equipment and turn them off. Then, it verifies if the aggregate bandwidth of the switches that had ports turned off is under a predefined threshold to determine whether their uplink ports speed should be reduced.

Finally, the PowerDatacenter class was extended by NetPowerDatacenter class to integrate the network model to CloudSim, allowing interaction with the events generated by other entities of the simulator. This class represents a data center with a network that comprises physical machines and access, aggregation and core switches; and computes its consumption during a simulation. Therefore, in each simulation frame, this class calculates the network power consumption, adds it to the data center's total power consumption and informs the state changes of servers to the network manager so it can reconfigure the network equipment according to the rules.

To perform the experiments, a main class, responsible for creating the scenario, starting the simulation and retrieving results, was implemented. It allows setting the characteristics of all the needed objects and the simulation parameters.

In order to facilitate simulations, a graphical interface that allows setting the scenario and repeating simulations with different parameters without modifying the source code was developed. This interface is showed in Figure 6.

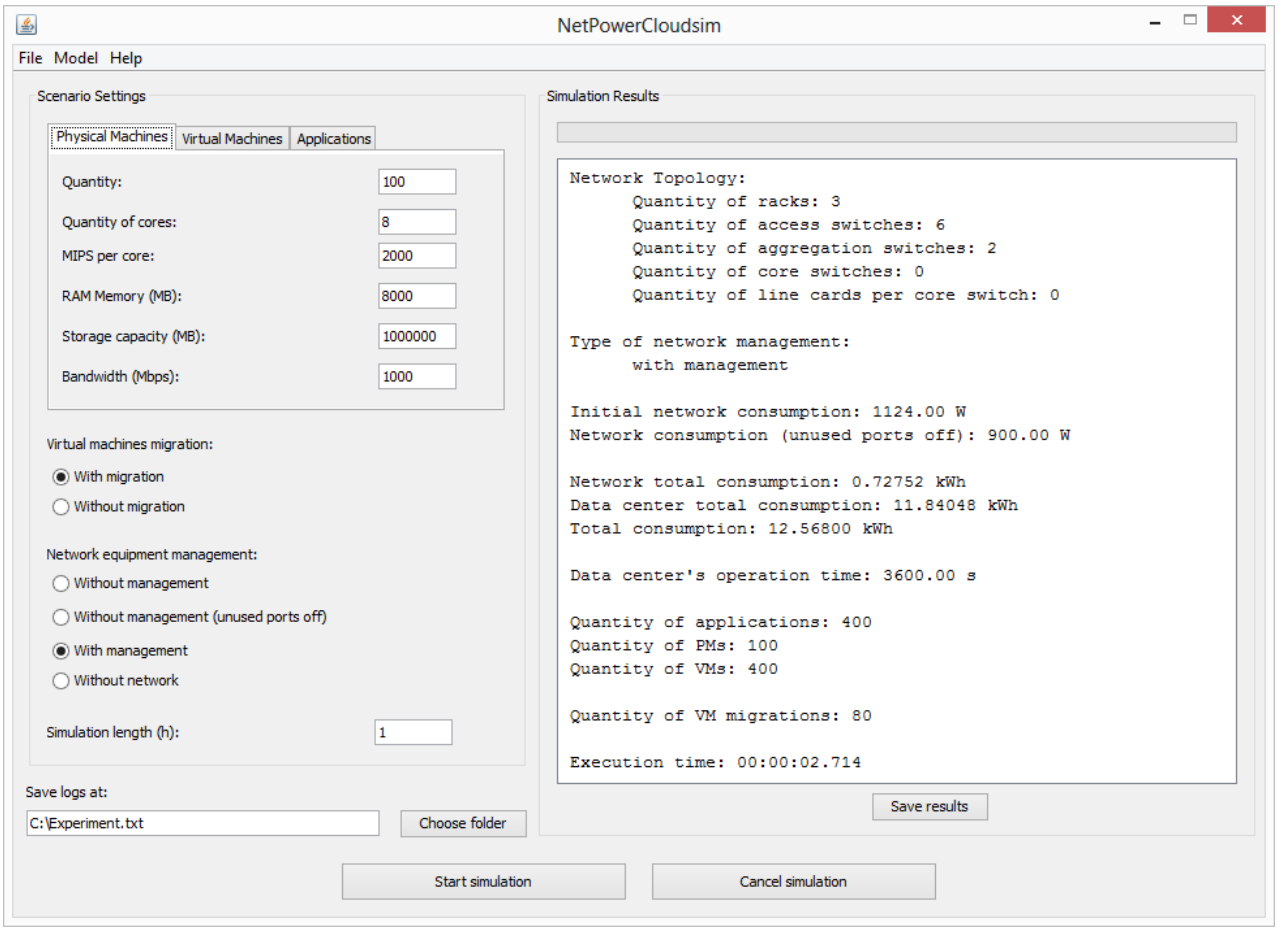


Figure 6. NetPowerCloudSim graphical user interface.

The version used for the extensions developing was CloudSim 3.0.3 and the implementation was made with the object oriented programming language Java through the development environment Netbeans IDE 7.3.1 along with the graphic library Swing. To verify each class code correction, unit tests were conducted using the JUnit framework. The integration tests were performed with the aid of a class developed for this purpose and by simple simulations that allowed thoroughly analyzing the logs and comparing the obtained results with those expected.

Lastly, for the sake of helping in understanding the code and the algorithms operation, as well as facilitating future extensions, all the created classes were widely commented and documentation was generated with the Javadoc tool, provided by Sun Microsystems.

B. *Experiments and Results*

In order to validate the model and the extensions, three experiments with different scenarios were conducted in a single machine. The experiments were simulations executed in a microcomputer with the following characteristics: Intel Core i5-3230M 2.6 GHz processor; 8 GB DDR3 RAM memory; and 64-bit Windows 8 operational system.

The scenarios created for the experiments had the following common parameters: 1 data center; physical machines with eight 2000 MIPS processing cores, 8 GB RAM memory, 1 TB storage capacity and 1 Gbps bandwidth; virtual machines with two 1000 MIPS processing cores, 1 GB RAM memory and 100 Mbps bandwidth.

The experiments performed and the results obtained are described and discussed next.

1) *Experiment 1*

In this experiment, one hour of operation of a data center with only 2 physical machines, 4 virtual machines and 4 applications was simulated, in order to verify the correct operation of the extensions and their interaction with CloudSim. The simulation was repeated four times and the results are presented in Table I: the first one was performed with the original version of CloudSim (R1); the second one, with NetPowerCloudSim without representing the network (R2); the third one, with network representation but without managing it (R3); and the last one, managing the network (R4). The results and the logs of each simulation were compared with each other and to the expected results and the necessary adjustments were made to ensure the correction and coherence of the model.

This experiment allowed evaluating all the functionalities of the developed extensions since, despite of the simplicity of the scenario, VMs migrations, PMs shutdowns and reductions and increases in PMs utilization rates happen and, consequently, network equipment speed configuration and ports shutdowns are performed.

In Table I it is possible to observe that, as expected, the data center servers consumption was constant over the four simulation repetitions and the network consumption was greater when it was not managed. It can also be observed that the network consumption has a very close value to the data center servers consumption. However, this happens because a rack was set up with two access switches for only two physical machines, which would not happen in a real infrastructure.

TABLE I. EXPERIMENT 1 RESULTS.

	Simulation repetition			
	***R*$_1$**	***R*$_2$**	***R*$_3$**	***R*$_4$**
Execution time (ms)	69	67	77	109
Network initial consumption (W*s)	-	-	188	124
Data center servers consumption (kWh)	0.1304	0.1304	0.1304	0.1304
Network consumption (kWh)	-	-	0.1723	0.1076
Total consumption (kWh)	0.1304	0.1304	0.3027	0.2380
MV migrations	2	2	2	2
PM shutdowns	1	1	1	1

2) *Experiment 2*

In this experiment, six hours of operation of a data center with 500 PMs was simulated, so that the network was composed by the three layers of the topology. In order to have a considerable amount of VMs migrations and PMs shutdowns, 2,000 VMs and 2,000 applications were executed. The simulation was repeated three times and the results obtained are shown in Table II: the first one, without managing the network (R1); the second one, also without managing the network, but turning off unused ports in the initial configuration (R2); and the last one, managing the network (R3).

It is possible to observe that the network consumption without management was 32.21 kWh; that, by turning off unused ports, there was a saving of 4.40 kWh (13.66%); and that, by managing the network equipment, the saving was increased to 6.85 kWh (21.25%). Considering the data center total energy consumption (401.62 kWh), there was a saving of 1.09% with unused ports off and 1.70% managing the network.

TABLE II. EXPERIMENT 2 RESULTS.

	Simulation repetition		
	***R*$_1$**	***R*$_2$**	***R*$_3$**
Execution time (min)	02:32.565	02:33.359	02:37.299
Network initial consumption (W*s)	5,444.00	4,700.00	4,700.00
Data center servers consumption (kWh)	369.4057	369.4057	369.4057
Network consumption (kWh)	32.2104	27.8084	25.3643
Total consumption (kWh)	401.6161	397.2141	394.7700
MV migrations	1,268	1,268	1,268
PM shutdowns	250	250	250

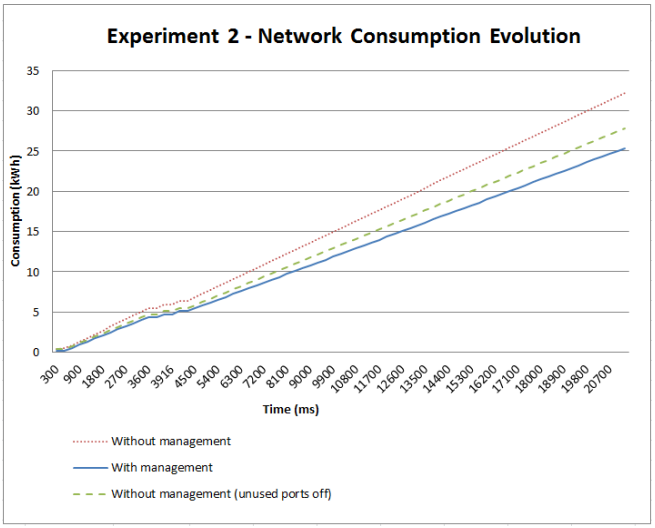


Figure 7. Experiment 2 network consumption evolution.

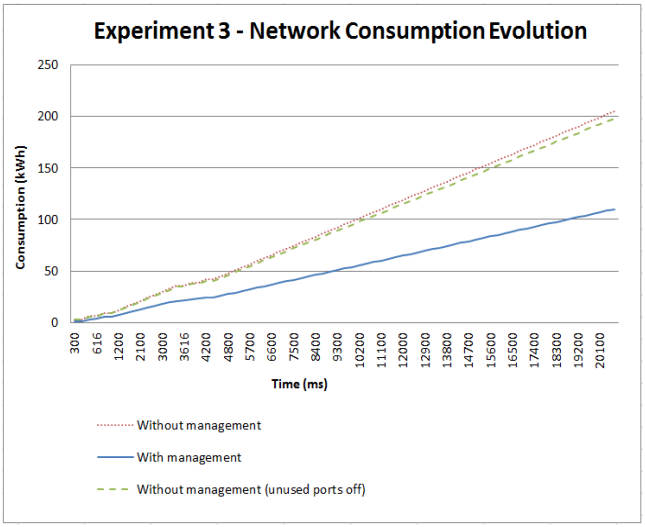


Figure 8. Experiment 3 network consumption evolution.

The chart in Figure 7 shows the evolution of the network energy consumption during the data center's 6-hour operation, representing the accumulated consumption in kWh at the end of each simulation frame.

3) Experiment 3

In this experiment, six hours of operation of a data center with 5,760 PMs, 10,000 VMs and 10,000 applications was simulated, with the purpose of testing the simulator representing a large scale data center and verifying the code's efficiency. As well as in Experiment 2, the simulation was repeated three times and the results are presented in Table III: the first one, without managing the network (R1); the second one, without managing the network, but turning off unused ports in the initial configuration (R2); and the last one, managing the network equipment (R3).

From the results of this experiment, it is possible to perceive that the network energy consumption without management was 211.06 kWh; that, by turning off unused ports in the initial configuration, there was a saving of 6.82 kWh (3.23%); and that, by managing the network equipment, the saving was increased to 97.92 kWh (46.39%). Considering the data center total consumption (2,071.77 kWh), there was a saving of 0.33% with unused ports off and 4.73% managing the network. The chart in Figure 8 shows the network energy consumption evolution during the data center's operation.

TABLE III. EXPERIMENT 3 RESULTS.

	Simulation repetition		
	R_1	R_2	R_3
Execution time (min)	24:49.597	25:28.382	27:09.126
Network initial consumption (W*s)	35,672.00	34,520.00	34,520.00
Data center servers consumption (kWh)	1,860.712	1,860.711	1,860.712
Network consumption (kWh)	211.060	204.244	113.141
Total consumption (kWh)	2,071.772	2,064.956	1,973.853
MV migrations	12,177	12,177	12,177
PM shutdowns	4,510	4,510	4,510

V. CONCLUSION AND FUTURE WORK

In this paper, basic concepts related to Green IT were first presented, i.e., Green Cloud and Green Networking, demonstrating the need of considering the network equipment in strategies designed to make data centers more efficient, since the network represents a significant percentage of total consumption, and this participation will be more expressive when the other components become more efficient.

Afterwards, in the related work section, a green cloud management model called OTM was presented, as well as network equipment management principles that, when properly applied, make the behavior of the total consumption of the network approximately proportional to the traffic load, even when legacy energy-agnostic equipment are used in. The proposal was to extend the OTM to manage the network traffic consolidation according to these management principles.

Then, the elements that must be added to the architecture of the OTM were described, including the rules and beliefs required for the correct network configuration according to the state changes of servers during the load consolidation process.

A model to simulate and validate the extensions to the OTM was also proposed. This model determines the data center network topology based on the number of physical servers, the rules to manage and set the network devices

according to the state changes of servers, and equations to calculate the switches consumption and the total network consumption.

The simulation model was validated by its application in a case study, which allowed verifying that equations and rules are correct and enough to create the topology and to calculate the consumption of the network in each step of the simulation, as well as highlight the possible effects of the application of the proposal. This model was the basis to create a simulator and perform simulations.

The simulator was created by extending CloudSim and was called NetPowerCloudSim. New classes to represent network equipment and network topology were created as well as a network manager that applies the rules during the simulation. A graphical interface was also developed in order to allow creating scenarios and perform simulation without the need to modify the application source code.

Finally, experiments to validate the extensions and the model were performed, demonstrating that it is possible to obtain significant energy savings in the data center consumption by the application of the model. It was thus demonstrated the possibility and desirability of extending the green cloud management model as proposed.

Although the actual results in each situation will depend on the data center configuration, the kind of network equipment and the workloads, it was demonstrated through the presented experiments that it is possible to obtain savings of nearly 50% in the network consumption and 5% in the data center total consumption in feasible conditions.

It is important to consider that the impact of applying the model is maximum in legacy energy-agnostic equipment, and will be smaller as the equipment becomes more energy-aware by applying the resources of the Green Networking but its application will be still convenient.

As future research, it is proposed to continue this work by performing experiments to determine the actual contribution of the model in scenarios with real configuration and workloads, as well as determine the most effective rules and virtual machine allocation policies. It is also proposed to compare these results to those obtained in real systems to calibrate the model.

Finally, the implementation of the model is proposed as future work and, since system performance may be affected if the network devices activation cost is bigger than the server activation cost, it is also suggested to study the proper network configuration and technologies to avoid this situation, with special consideration to protocols that manage the links redundancy and aggregation, like the Spanning Tree Protocol, MC-LAG, and other new networking standards for data centers.

REFERENCES

[1] S. R. Villarreal, C. B. Westphall, and C. M. Westphall, "Optimizing green clouds through legacy network infrastructure management," Proc. Thirteenth International Conference on Networks (ICN – 2014), IARIA XPS Press, 2014, pp. 142-147.

[2] C. B. Westphall and S. R. Villarreal, "Principles and trends in Green Cloud Computing," Revista Eletrônica de Sistemas de Informação, vol. 12, n. 1, pp. 1-19, January 2013, doi: 10.5329/RESI.2013.1201007.

[3] A. Beloglazov, R. Buyya, Y. C. Lee, and A. Zomaya, "A taxonomy and survey of energy-efficient datacenters and Cloud Computing," Advances in Computers, vol. 82, pp. 47-111, Elsevier, November 2011, doi: 10.1016/B978-0-12-385512-1.00003-7.

[4] R. Calheiros, R. Ranjan, A. Beloglazov, C. De Rose, and R. Buyya, "CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resources provisioning algorithms," SPE Wiley Press, vol. 41, January 2011, pp. 23-50, doi: 10.1002/spe.995.

[5] A. Abdullah, "Green Cloud Computing: the need of the hour," International Journal of Research in Advent Technology, vol. 2, n. 1, January 2014, pp. 316-321.

[6] C. B. Westphall, C. M. Westphall, S. R. Villarreal, G. A. Geronimo, J. Werner, Green Clouds through Servers, Virtual Machines and Network Infrastructure Management. In book: Courses / 32nd Brazilian Symposium on Computer Networks and Distributed Systems, Chapter: 6, Publisher: SBC – SBRC 2014, vol. 1, pp. 244-289.

[7] A. Bianzino, C. Chaudet, D. Rossi, and J. Rougier, "A survey of green networking research," IEEE Communications Surveys and Tutorials, vol. 14, pp. 3-20, February 2012, doi: 10.1109/SURV.2011.113010.00106.

[8] S. Jing et al, "State-of-the-art research study for Green Cloud Computing," The Journal of Supercomputing, vol. 65, n. 1, July 2013, pp. 445-468.

[9] P. Mahadevan, P. Sharma, S. Banerjee and P. Ranganathan, "A power benchmarking framework for network devices," Proc. 8th International IFIP-TC 6 Networking Conference, Springer Berlin Heidelberg, November 2009, pp. 795-808, doi: 10.1007/978-3-642-01399-7_62.

[10] P. Mahadevan, S. Banerjee, P. Sharma, A. Shah, and P. Ranganathan, "On energy efficiency for enterprise and data center networks," IEEE Communication Magazine. vol. 49 pp. 94-100. August 2011. 10.1109/MCOM.2011.5978421.

[11] J. Werner, "A virtual machines allocation approach in green cloud computing environments". Dissertation: Post-Graduate Program in Computer Science Federal University of Santa Catarina, 2011.

[12] R. Freitas, "Efficient energy use for cloud computing through simulations". Monograph: Post-Graduate Program in Computer Science Federal University of Santa Catarina, 2011.

[13] S. Garg and R. Buyya, "NetworkCloudSim: modelling parallel applications in cloud simulations," Fourth IEEE International Conference on Utility and Cloud Computing, IEEE, December 2011, pp. 105-113, doi: 10.1109/UCC.2011.24.

[14] A. Beloglazov, "Energy-efficient management of virtual machines in data centers for Cloud Computing". PhD Thesis, University of Melbourne, Australia, 2013.

[15] C. Sher De Cusatis, A. Carranza, and C. Decusatis, "Comunication within clouds: open standards and proprietary protocols for data center networking," IEEE Communication Magazine, vol. 50, pp. 26-33, September 2012. doi: 10.1109/MCOM.2012.6295708.

[16] J. Werner, G. Geronimo, C. B. Westphall, F. Koch, and R. Freitas, "Simulator improvements to validate the green cloud computing approach," LANOMS, October 2011, pp. 1-8, doi: 10.1109/LANOMS.2011.6102263.

[17] K. Christensen, P. Reviriego, B. Nordman, M. Mostowfi, and J. Maestro, "IEEE 802.3az: The road to energy efficient Ethernet," IEEE Comunication Magazine, vol. 48, pp. 50-56, November 2010. doi: 10.1109/MCOM.2010.5621967.

From Multi-disciplinary Knowledge Objects to Universal Knowledge Dimensions: Creating Computational Views

Claus-Peter Rückemann
Westfälische Wilhelms-Universität Münster (WWU),
Leibniz Universität Hannover,
North-German Supercomputing Alliance (HLRN), Germany
Email: ruckema@uni-muenster.de

***Abstract*—Creating and deploying long-term knowledge resources requires research on content as well as on application side. This paper presents the results on high-end structure and classification, being the base for an improved long-term documentation and providing advanced computational views. Content and application context cover factual, conceptual, procedural, and metacognitive knowledge. The core elements are multi-disciplinary knowledge objects, which carry attributes allowing for the creation of various dimensions inside the knowledge resources. The attributes are accompanied by references to flexible multi-lingual universal classifications. The discussion presents the fundamental results having contributed to the knowledge resources. The organisational structure of the knowledge resources supports Big Data integration by a sustainable knowledge definition. The outcome are multi-disciplinary knowledge bases linked with any context, which have led to generate knowledge matrices and which can be used to create computational views. The case studies implemented content and structure as well as workflows for creating knowledge matrices. The implementation deploys dynamical, interactive, and batch computing and storage resources in an Integrated Information and Computing System environment exploiting High End Computing (HEC) and High Performance Computing (HPC) resources and references to the Universal Decimal Classification (UDC). The paper discusses the new practical results from the accompanying long-term projects and implementation.**

Keywords–Knowledge Resources; Computational views; Multi-disciplinary documentation; Advanced Computing; Sustainability; Information Systems; Classification; UDC; Natural Sciences.

I. INTRODUCTION

This article is build on the work performed on advanced computing for the processing of complex knowledge-based information [1]. It has been shown that knowledge processing considerably benefits from employing universal classification.

The demand for long-term sustainability of the resources increases with the complexity of content and context. The organisation and structure of the resources are getting essentially important, the more important the more the data sizes and complexity as well as their intelligent use are required [2]. The article therefore introduces and discusses the background, a comprehensive knowledge definition, including the systematics and methodologies required for an advanced long-term documentation, which can be deployed in most flexible ways. The general requirements have to consider the condition that it is not sufficient to support only an isolated or special methodology. The knowledge requires special qualities in order to be usable as well as the quantities of knowledge counts. A suitable general conceptual handling and a universal knowledge definition is required in this environment for supporting a high quality of resulting context and matrices.

The results are the outcome of the case studies conducted over the last years. The article discusses the fundamental components developed and employed for enabling a systematical processing for arbitrary disciplines. It presents examples for complementary knowledge and sectional views, which can be considered from the knowledge dimensions. Examples for generator workflows and considerations on statistics are the base for several case studies in geosciences and archaeology knowledge. The main new contribution of this research is the implementation of using a universal classification for multi-disciplinary and multi-lingual knowledge classification, for example, for dynamical and computational views.

This paper is organised as follows. Section II introduces with the qualities and quantities, especially with the implemented knowledge resources and organisational structures, data integration, and the concrete practical knowledge definition used. In this context, this section shows how others address challenges of classification, knowledge, and Big Data. Section III discusses the previous work and components employed, Section IV presents the systematics and classification used for the processing, Section V shows the conceptual base for the knowledge dimensions and examples for practical section views. Section VI introduces in the resulting matrix generators and computational consequences. Section VII delivers a discussion of the results from the extended implementation case study. Section VIII shows the results from an external knowledge integration from web resources. Sections IX and X evaluate main results, and summarise the lessons learned, conclusions and future work.

II. KNOWLEDGE QUALITIES AND QUANTITIES

The essential base for any implementation is the creation of suitable content comprising all aspects of knowledge. We have to discuss the consequences of a systematical use, the data itself as well as the basic definition resulting from this work and providing the fundaments for future developments.

A. Knowledge and organisational structures

The systematical use of knowledge resources has led to organising the resources in structures allowing objects and

containers. The application of conceptual knowledge on these elements has led to the creation of knowledge dimensions, which can be used with references to the universal classification.

A practical application is the generation of computational views. Example cases studies presented for this extended research are from two different disciplines, volcanological features and meteorite impact structures, whose attributes and features provide criteria for classification, indicators, and multi-disciplinary research. Adding structures, classification, and other qualities to large knowledge resources implicitly increases the computational and storage requirements for the factual knowledge resources. But at the same time the additional conceptual, procedural, and metacognitive knowledge drastically contributes to the optimisation of workflows, to the documentation and integration of knowledge resources, and the quality of applications and results.

These knowledge dimensions effectively reduce the computational and storage requirements for complex and integrated knowledge, supporting really Big Data scenarios. At the same time the requirements for the conceptual knowledge services, the classification, are high regarding width and depth.

For the referenced classification, which has shown up being most important with complex multi-disciplinary long-term classification with practical simple and advanced applications of knowledge resources is the Universal Decimal Classification (UDC) [3]. According to Wikipedia currently about 150,000 institutions, mostly libraries and institutions handling large amounts of data and information, e.g., the ETH Library (Eidgenössische Technische Hochschule), are using basic UDC classification worldwide [4], e.g., with documentation of their resources, library content, bibliographic purposes on publications and references, for digital and realia objects. Regarding the library applications only, UDC is used in more than 144,000 institutions and 130 countries [5]. Further operational areas are author-side content classifications and museum collections.

UDC allows an efficient and effective processing of knowledge data. UDC provides facilities to obtain a universal and systematical view on the classified objects. UDC in combination with statistical methods can be used for analysing knowledge data for many purposes and in a multitude of ways.

With the knowledge resources in this research handling 70,000 classes, for 100,000 objects and several millions of referenced data then simple workflows can be linear but the more complex the algorithms get the workflows will mostly become non-linear. The workflows allow interactive use, dynamical communication, computing, decision support, management, and pre- and postprocessing, e.g., visualisation.

Besides the content there are reasons from the application side, which introduce non-linear behaviour. If going into real application cases, as the ones presented here, the linearity depends on the workflow and data. Any real workflow is subject of limitation, e.g., time, resources, generation of excerpts, and amount of ergonomically practical matrix elements. Creating and exploiting growing knowledge resources regulary goes far beyond the growth of many end-user scenarios. Nevertheless, the growing resources provide excellent means for improving workflows end results. Therefore, the source matrices and the number of references grow much faster than the result matrices, which inevitably make these processes non-linear.

The meaning of knowledge management changed with the extended use, especially when covering the development from "information society" to "knowledge society". The reasons for the change are resulting from advances in quality, efficiency, precision, consistency and a systematical long-term use, which in combination can contribute reducing digital gaps. In consequence this also promotes new technologies with used resources, e.g., in knowledge management [6] and Library Information Systems (LIS) [7]. The advances in creating new resources pushes the features for integration of resources and for the development of advanced views.

B. Knowledge and Big Data

Many critical reflections on Big Data activities [8] argue that most of the data [9] is unstructured, which is a central issue against an efficient handling of the content. On the other hand, examples of the fundamental aspects of classification in traditional library-focussed context show large benefits of applying even a small extend of knowledge in a well defined environment [10]. The multi-disciplinary aspects from science, economy, society [11] as well as the attribution of "Big Wealth" [12] have triggered new perspectives.

On the one hand, new visions in research [13] and education [14] focus on intelligent and smart environments and components [15]. Not surprisingly, the industry view on the optimisation of "cognition" from Big Data is more or less short term and application centric [16].

On the other hand, the wealth of data implies consequences for anyone handling data and developing applications, which especially results in increased potential [17] and also in challenges [18]. with technologies, methodologies, and systematics.

The overview reveals that there are differences when working with "Big Data" for various academic research on the one hand and use in industry and economy on the other hand. Besides science and industry, assessing knowledge loss risks [19] resulting from departing personnel [20] can be summarised by the risk of knowledge loss, the probability for loss of employees, the consequences of human knowledge loss, and the quality of knowledge resources.

Quality and risks of knowledge loss are correlated with the assessment of management positions [21]. Especially the desintegration of knowledge is hand in hand with the desintegration of workflow and system components. The importance of external auditing for the casting of management and decisions increases with the size of centers and with mission critical services. Exactly the knowledge loss with Big Data then means "big loss of knowledge", especially as the loss in mostly quality, which can even be worse as the quantity of less quality data is increasing at the same time.

The overall big data challenges, data intensive volume, variability, velocity and for future scenarios especially data vitality,

meaning long-term documentation, usability, and accessibility can be handled in a scalable, modular way. The often publicly proposed three "V" for Big Data, namely volume, variability, velocity have been exended by volatility, and veracity in the last years. All of these a primarily describing technical, non-content related issues. For long-term relevant Big Data, especially if the share in structure is significant and even increasing, then the vitality is of critical value.

The approach for coping with Big Data starts with the application of conceptual knowledge. Hence, conceptual knowledge as with an implementation on universal classification provides a "Knowledge as an Infrastructure" solution.

Practical examples are knowledge integration for scientific classification and computation [22], flexible general object envelopes [23], and the management of knowledge and resources, e.g., for environmental information and computation, which has become a focus application [24] in environmental and disaster management. Research conducted on environmental protection and climate change, integrating multi-disciplinary and multi-lingual knowledge resources [25] on an international and trans-national base has shown that sustainable collaboration and governance requires long-term knowledge, classification, and standards as well as management system components.

C. Knowledge definition and understanding

The World Social Science Report 2013 [26] defines knowledge as "The way society and individuals apply meaning to experience ...". Accordingly, the report proposes that "New media and new forms of public participation and greater access to information, are crucial" for open knowledge systems.

In general, we can have an understanding, where knowledge is: Knowledge is created from a subjective combination of different attainments as there are intuition, experience, information, education, decision, power of persuasion and so on, which are selected, compared and balanced against each other, which are transformed and interpreted.

The consequences are: Authentic knowledge therefore does not exist, it always has to be enlived again. Knowledge must not be confused with information or data, which can be stored. Knowledge cannot be stored nor can it simply exist, neither in the Internet, nor in computers, databases, programs or books. Therefore, the demands for knowledge resources in support of the knowledge creation process are complex and multifold.

There is no universal "definition" of the term "knowledge", but UDC provides a good overview of the possible width, depth, and facets. For this research the classification references of UDC:0 (Science and knowledge) define the view on universal knowledge.

1) Big Data: Knowledge Top Level: The question "What is knowledge?" in the conceptual knowledge dimension is best answered by the appropriate classification used with the knowledge resources' application scenarios (Table I). For this case the table shows an excerpt of the knowledge top level classification (Universal Decimal Classification, UDC) used with the knowledge resources.

TABLE I. UNIVERSAL DECIMAL CLASSIFICATION: KNOWLEDGE TOP LEVEL CLASSIFICATION WITH KNOWLEDGE RESOURCES.

UDC Code	*Description (English, excerpt)*
UDC:0	Science and knowledge. Organization. Computer science. ...
UDC:00	Prolegomena. Fundamentals of knowledge and culture ...
UDC:001	Science and knowledge in general. Organization ...
UDC:002	Documentation. Books. Writings. Authorship
UDC:003	Writing systems and scripts
UDC:004	Computer science and technology. Computing
UDC:004.2	Computer architecture
UDC:004.3	Computer hardware
UDC:004.4	Software
UDC:004.5	Human-computer interaction
UDC:004.6	Data
UDC:004.7	Computer communication
UDC:004.8	Artificial intelligence
UDC:004.9	Application-oriented computer-based techniques
UDC:005	Management
UDC:005.1	Management Theory
UDC:005.2	Management agents. Mechanisms. Measures
UDC:005.3	Management activities
UDC:005.5	Management operations. Direction
UDC:005.6	Quality management. Total quality management (TQM)

This classification reflects the conceptual dimension and is intended to be used with the full bandwidth of knowledge and knowledge resources.

2) Knowledge processing and application: Geoscientific knowledge processing is traditionally focussed on processing and analysis of data resulting from geophysical or geological measurements. Examples are processing based on seismological, seismic, magnetic, or gravimetric data. The amount of information and documentation from geosciences and natural sciences based methods and features as well as their complexity has steadily increased for decades. Efficiency and economical practice forces to long-term document and exploit this pool of multi-disciplinary information. Spatial and chronological data and classification are an indispensable component. It is becoming increasingly important that with most professional analysis different geophysical methods and results have to be used in combination.

Common means of application and knowledge discovery, e.g., isolated batch or interactive application scenarios or string based search routines on plain data cannot even approximately integrate the required higher complexity of real environments.

The knowledge gathered during generations should be considered the most valuable component, the more important for long-term results from geosciences. The universal knowledge resources require long-term documentation as well as universal classification and structuring, beyond traditional collections [27], digital libraries [28] and isolated content. With the long-term multi-disciplinary resources the high end processing and computing aspects are essential for sustainability and discovery. Therefore, it is recommended to implement scientific supercomputing resources supporting advanced information systems and creating and improving workflows as recommended [29] with Integrated Information and Computing System (IICS) components [30] and High End Computing (HEC) [31]. This paper presents the results from creating

and managing long-term knowledge resources for knowledge processing by employing a universal classification like the Universal Decimal Classification (UDC) [3]. It discusses the experiences handling systematics and classification as well as the methodological use of "Object Carousels". The paper points out the demands and challenges as resulting from the case studies within the GEXI collaborations [32] concentrating on integrating knowledge from geosciences, volcanology, and spatial sciences disciplines.

III. Components employed

The data used here is based on the content and context from the LX Foundation Scientific Resources [33] and corresponding case studies [34]. The LX structure and UDC [3] with its features [35] are an essential means for the processing workflows and evaluation of the knowledge objects and containers. The applied workflows and processing are based on the data and extended features developed for the Gottfried Wilhelm Leibniz resources [36]. Although shown in detail with previous publications, the following terms may be useful when discussing the knowledge resources and application components.

- Object: An entity of knowledge data being part of knowledge resources. An object can contain any documentation, references, and other data. Objects can have an arbitrary number of sub-objects. Example: Description and location of an archaeological site, locations being part of the location may be handled as sub-objects.
- Container: A collection of knowledge objects in a conjoint format. Example: Volcanological features database.
- Matrix: A subset of the entirety, the "universe", of knowledge. A workflow can consist of many subworkflows each of which can be based on an arbitrary number of knowledge matrices. The output of any subworkflow or workflow can be seen as an intermediate or final result matrix. Example: The output elements of a discovery or search request.
- Qualities: The entirety of documentation including attributes and data being part of knowledge objects.
- Quantities: The number of objects available.
- Systematics: The systematics, a plan based strategy, used for creating knowledge resources for disciplines as well as the systematical use of knowledge resources, e.g., with conceptual knowledge.

The classification, the result of assigning and arranging in classes, is state-of-the-art within the knowledge resources, which implicitly means that the classification is not created statically or even fixed. It can be used and dynamically be modified on the fly, e.g., when required by a knowledge discovery workflow description. Representations and references can be handled dynamically with the context of a discovery process. So, the classification can be dynamically modelled with the context of the workflow.

The context is made up from all the properties and qualities, which have been added to a resource. These properties and qualities can be references, text, bibliographic data, content of publications, links, classification, keywords, factual, conceptual or procedural knowledge and so on.

The granularity must be floating, it depends on the efforts a research community wants to invest for a specific documentation and application. This repeatedly depends on the type of data, the purpose, the actuality and so on. Any of these objects can be evaluated. It depends on the decision of a group or service, which criteria or automatism to take into account.

The LX resources can provide any knowledge documentation and additional information on objects as well as, e.g., geo- and knowledge references. The volcanological data used in the examples is embedded into millions of multi-disciplinary objects, dynamical and spatial information and data files.

The knowledge objects are under continuous development for more than twenty-five years. The classification information has been added in order to describe the objects with the ongoing research and in order to enable more detailed documentation in a multi-disciplinary context. The knowledge resources can make sustainable and vital use of Object Carousels [37] in order to create knowledge object references and to modularise the required algorithms [38]. This provides a universal means for improving coverage, e.g., dark data, and quality within the workflow.

This research focusses on the organisation of knowledge resources and computational views. The general aspects of components and algorithms are focus of different research studies as they mostly belong into the task of the different services and disciplines. The architecture of the components for the purpose of advanced scientific computing and multidisciplinary documentation is described (starting with Figure 1) in [39]. An example for a workflow, algorithm, and result matrix scenario on knowledge integration for scientific classification and computation is given in [22].

Therefore, for the cases presented in this research paper we had to concentrate on the structure of the knowledge objects, georeferencing, and references data. As secondary components, besides IICS applications and interfaces are available, allowing parallel workflows [40] and intelligent components [41] on HEC and HPC resources. With the IICS, the Generic Mapping Tools (GMT) [42] have been used to visualise georeferenced data wherever a spatial representation is reasonable.

IV. Disciplines, systematics, and processing

In geosciences, there is no globally unique stratigraphy. Different continents and regions require different and detailed stratigraphies. Therefore, it is not practicable to have a flat unique global standard due to the regional differences in geological development. Present common stratigraphy concepts [43] fail on general use as well as on a consistent universal classification required. Implementing a universal long-term use we further need to consider appropriate systematics, e.g., lithostratigraphical, chronological, biostratigraphical, chronometrical, chronostratigraphical systematics. For example when it comes to plants, animals, and genotype "-zoic", "-phytic" or "-gen" often mix without distinction. This is the case with many languages, for example, "Mesozoic / Mesophytic" respective "Mesozoikum / Mesophytikum".

Instead of a mix-up of terminology, for systematical use the alignment of the Eonothems / Eons, Erathems / Eras, Systems / Periods, Series / Epochs, and Stages / Ages and so on should be handled consistently and consequently. In addition, the multi-regional dimension should be available for these, showing correspondence with the appropriate absolute ages, as available on-site.

For an efficient and effective processing the knowledge data requires a flexible structure and a universal systematic classification. Any knowledge resources documenting complex multi-disciplinary reality for discovery applications require features for exact documentation on the one hand and they require soft criteria on the other hand.

UDC is a classification complying with the classification criteria. Together with the content, which may deliver more detail or differing perspectives UDC provides a universal view on the classified objects. When requiring facetted classification for multi-disciplinary knowledge the universal UDC cannot be ignored as it is the most comprehensive and flexible means available and supported.

The classification deployed for documentation [44] is able to document any object with any relation, structure, and level of detail as well as intelligently selected nearby hits and references. Objects include any media, textual documents, illustrations, photos, maps, videos, sound recordings, as well as realia, physical objects such as museum objects. UDC is a suitable background classification, for example:

The objects use preliminary classifications for multi-disciplinary content. Standardised operations used with UDC are coordination and addition ("+"), consecutive extension ("/"), relation (":"), order-fixing ("::"), subgrouping ("[]"), non-UDC notation ("*"), alphabetic extension ("A-Z"), besides place, time, nationality, language, form, and characteristics.

V. Dimensions and creating computational views

The dimensions available mostly depend on the features and complexity of the available data. Therefore, a number of essential aspects have been considered when creating content with the knowledge resources. Regarding the views many new arrangements and visualisations are possible.

Nevertheless, it can be quite challenging for application developers to create representations, which can be visualised and to implement suitable components. In many cases so called "Section Views" can be computed, which use n-dimensional sections from 'n+m'-dimensional knowledge context.

A. Implementation of knowledge dimensions

The implemented knowledge resources integrate factual, conceptual, procedural, and metacognitive knowledge. Table II shows the major types of knowledge as complementary parts of the knowledge resources. The table shows some practical examples, which illustrate the benefits of the integration. The data itself is represented in knowledge objects containing any kind of information and collections, including content, classifications, and references.

Table II. Complementary knowledge in the knowledge resources, types and examples.

Knowledge	*Examples*
Factual knowledge	Terminologies
	Factual details
Conceptual knowledge	Classifications, categorisations
	Principles, generalisations
	Theories, models, structures
Procedural knowledge	Algorithms, workflows, skills
	Methods, techniques
	Determination on procedures, decision making
Metacognitive knowledge	Strategies
	Self-knowledge
	Cognitive tasks, contexts, conditions

With the content and context documentation the knowledge objects describe the integrated knowledge space, for which any dimensions can be interconnected.

B. Section views

The implemented structure and content enable to create section views based on the knowledge dimensions, which can, e.g., be physical or contextual dimensions. Table III shows some section views, which can be based on the combination of contextual and factual knowledge.

Table III. Section views based on the knowledge dimensions. Section views and examples in practice.

Section Views	*Examples*
Attributes	colour, size, ...
	extremes ...
Space and location	spatial distributions
	geo-references
	depth distribution
Time	timelines
	time index
Cultures	context
	history
	time
	location
	society
	inventions
	art
Disciplines	physics
	geophysics
	archaeology
Multi-disciplines	geosciences
	natural sciences – humanities
Multi-lingual	English
	German
	Romanic language
Combinations	depth distribution - timelines
	location-fixed: Objects over time
	time-fixed: Objects over space/locations
	culture-fixed: Objects over space and time

Generators can access the knowledge resources and their workflows can apply any appropriate components and algorithms, e.g., classifications, phonetics [45], associations, references, keywords, and statistics.

VI. GENERATORS BASED ON THE DIMENSIONS

The knowledge resources are the base for extended modular matrix generators. Examples are report generators. Nevertheless, report generators can become arbitrarily complex, they are only simple compared with matrix generators in complex environments. A generator can access any data from any requested knowledge resources available. An implementation can integrate different sources as well as applications for various purposes. An example for a simple standard generator can be summarised as

1) Workflow description,
2) Knowledge resources interface,
3) Context selection,
4) Regular expressions,
5) Matrix generation,
6) Result ⇒ Term-based matrix.

The knowledge dimensions are considered in 1), their resources are chosen in 2), the context of the dimensions is selected in 3), and the objects are selected in 4). A simple advanced generator can integrate external core interfaces for extended capabilities:

1) Workflow description,
2) Knowledge resources interface,
3) Context selection,
4) Container interface,
5) Section view selection,
6) LaTeX core interface,
7) Makeindex generation,
8) lxchaidx and further special configuration,
9) Index formatting / index style,
10) Result ⇒ Index-structured matrix.

This example workflow integrates a core interface for the creation of index-structured result matrices. The result matrices contain cross referenced objects integrated from the resources and the selected containers.

The structure of the content is important for the interfaces and for the generation of the results. Besides other factors, the structure and organisation of the content is responsible for the universal deployment of features and therefore an important factor for the quality of result matrices. The workflows can integrate plausibility checks depending on the scenario. These include consideration of structure, content, and knowledge.

Currently, one Central Processing Unit (CPU) can handle sizes of reasonably comprehensive atomic context. A 100,000 object base requires about 1 hour per run per index, including formatting and references. For interactive and dynamical applications where no huge index-structured matrices may be required the object base for an atomic context can be about 100 objects. Any of these scenarios can require parallel instances on available resources instead of consecutive runs before the matrix or index generation is done.

The data herewith can be extended with these steps referencing different sources, creating synopsis matrices, collecting registers, and processing of source data.

Proceeding this way, long-term benefits can accumulate by a sustainable re-use of referenced and source data. Generators based on core knowledge resources, which are being evolved can deliver results-at-a-time. This means that the results can depend on the actual state of the resources (content, context, references and so on).

For a re-use, these results-at-a-time can be preserved. It depends on the scenario, which context data have to be kept for intelligent reports based on the preserved results. For example, the results might be usable with the original sources or snapshot of the objects so that these might be reasonable to be preserved, too.

If knowledge resources are extended this regulary goes ahead with the wish to increase the computing capacities. Contributing conceptual knowledge delivers essential benefits for the knowledge resources but it is also an add-on to the overall data sizes and requirements [46]. This is a classical case for using High End Computing resources, improving results by deploying advanced complex knowledge resources and still reducing computing times.

Even more important is the fact that the use of classification can support the creation of very scalable solutions. The reason for this is that rigid structures are replaced by a matrix of flexible and elastic objects allowing references and a multitude of methods.

The main reasons for the implementation is that the knowledge resources allow

- flexible and dynamical matrix formation, e.g., groups and subgroups and
- to displace the implementation focus from a relatively primitive increase of computational power into the direction of the improvement of structures and workflows.

Example: Adding some of the simplest classification (discipline A and discipline B) to 500 knowledge objects of discipline A and B each adds 1,000 classification entries but at the same time searching for A reduces the number of objects to be handled with their content by half. Even in a slightly more complex workflow the efficiency improvement can easily be in the range of thousands of percents.

Computing result matrices is an arbitrary complex task, which can depend on various factors. Applying statistics and classification to knowledge resources has successfully provided excellent solutions, which can be used for optimising result matrices in context of natural sciences, e.g., geosciences, archaeology, volcanology or with spatial disciplines, as well as for universal knowledge.

The method and application types used for optimisation imply some general characteristics when putting discovery workflows into practice regarding components like terms, media, and other context (Table IV). Regarding the computational views the table lists some representative numbers for Section Views (SV) in addition.

TABLE IV. RESULTING PER-INSTANCE-CALLS FOR METHOD AND APPLICATION TYPES ON OPTIMISATION WITH KNOWLEDGE DISCOVERY.

Type	*Terms*	*Media*	*Workflow*	*Algorithm*	*Combination*
Mean	500	20	20	50,000	3,000
Median	10	5	2	5,000	50
Deviation	30	5	5	200	20
Distribution	90	40	15	20	120
Correlation	15	10	5	20	90
Probability	140	15	20	50	150
Phonetics	50	5	10	20	50
Regular expr.	920	100	50	40	1,500
References	720	120	30	5	900
Association	610	60	10	5	420
UDC	530	120	20	5	660
Keywords	820	100	10	5	600
Translations	245	20	5	5	650
Corrections	60	10	5	5	150
External res.	40	30	5	5	40
SV time	1,100	25	6,000	15	2,400
SV space	850	10	2,500	15	1,800
SV attribute	20	4	70	15	1,200
SV discipline	55	8	5	15	1,500
SV culture	5	4	5	15	760

Statistics methods have shown to be an important means for successfully optimising result matrices. The most widely implemented methods for the creation of result matrices are intermediate result matrices based on regular expressions and intermediate result matrices based on combined regular expressions, classification, and statistics, giving their numbers special weight.

Based on these per-instance numbers this results in demanding requirements for complex applications –

- On numerical data: Millions of calls are done per algorithm and dataset, hundreds in parallel/compact numeric routines.
- On "terms": Hundred thousands of calls are done per sub-workflow, thousands in parallel/complex routines, are done.

Most resources are created for one application scenario or are used for one application scenario only. Only 5–10 percent overlap between disciplines – due to mostly isolated use. Large benefits result from multi-disciplinary multi-lingual integration, which is a major contribution enabling to create SV from resources.

The multi-lingual application adds an additional dimension to the knowledge matrix, which can be used by most discovery processes. As this implemented dimension is of very high quality the matrix space can benefit vastly from content and references.

Section Views show a section through any of the dimensions in the knowledge resources. However, a limited number of possible views is commonly used. The table lists some more regulary used SV based on time, space, attribute, discipline, and culture. Section View reduces the dimensions to a number that can be further used, for example, analysed or visualised.

The interesting aspect is that the means of time and space are most widely used in workflows and combined methods and they are mostly associated with terms, which implies the fact that terms are handled with means like regular expressions. Other types like attributes, discipline, and culture are more or less represented, which results from the fact that the awareness of this context is still not widely spread.

VII. CASE STUDY IMPLEMENTATION AND RESULTS

The following sections discuss the work done for using knowledge resource objects with processing and computing from within IICS. For knowledge resources it is necessary that any classification can be added while the content is developed, over long period of time, more than decades. With the cases presented the content has been created over more than twenty-five years.

- Methodologically, in a first phase, objects have been documented without classification.
- In a second phase, all objects describing volcanic features have been classified as volcanic features.
- In a third phase, volcanic features' objects have been classified into separate classes as required with ongoing extended description of objects in a multi-disciplinary context.

The case study presents a state-of-the-art selection of volcanological and geological features. An evaluation of the association that users have, showed that the criteria "date" and "location" are most prominent with objects if the workflow approaches from the "surface (of the earth)" view [32]. Mapping and timelining with all the respective views will be the natural result.

The small unsorted excerpts of the knowledge resources objects only refer to main UDC-based classes, which for this part of the publication are taken from the Multilingual Universal Decimal Classification Summary (UDCC Publication No. 088) [47] released by the UDC Consortium under the Creative Commons Attribution Share Alike 3.0 license [48] (first release 2009, subsequent update 2012).

As with any object, algorithms, phonetics, regular expressions, statistics, complementary translation or transliteration objects, and other features and discovery processes can be combined for facets and views for any classification subject. Complementary, objects on arbitary algorithms, e.g., processing or statistics, can be included in the knowledge resources and provisioned and applied for further use.

1) Time: Table V shows an excerpt of the resulting UDC classification of spaces of time practically used with the knowledge resources. Instead of the earlier UDC editions the classifications are composite UDC:551.7 mappings referring to historical geology and stratigraphy for all the spaces of time.

TABLE V. UNIVERSAL DECIMAL CLASSIFICATION OF SPACES OF TIME USED WITH THE KNOWLEDGE RESOURCES (EXCERPT).

UDC Code	*Description (English)*
UDC:"0/2"	Dates and ranges of time (CE or AD) ...
UDC:"0"	First millennium CE
UDC:"1"	Second millennium CE
UDC:"2"	Third millennium CE
UDC:"3/7"	Time divisions other than dates in Christian ...
UDC:"3"	Conventional time divisions and subdivisions ...
UDC:"4"	Duration. Time-span. Period. Term. Ages ...
UDC:"5"	Periodicity. Frequency. Recurrence at ...
UDC:"6"	Geological, archaeological and cultural time divisions
UDC:"61/62"	Geological time division
UDC:"63"	Archaeological, prehistoric, protohistoric periods ...
UDC:"67/69"	Time reckonings: universal, secular, non-Christian ...
UDC:"67"	Universal time reckoning. Before Present
UDC:"68"	Secular time reckonings other than universal and ...
UDC:"69"	Dates and time units in non-Christian ...
UDC:"7"	Phenomena in time. Phenomenology of time
UDC:551.7+"61"	Cryptozoic aeon. Precambrian. 600+ MYBP ...
UDC:551.7+"616"	Archaean. Ur-gneiss formation. Ur-schiefer formation
UDC:551.7+"618"	Eozoic. Algonkian
UDC:551.7+"62"	Phanerozoic aeon. 600 MYBP - Present
UDC:551.7+"621"	Palaeozoic. 600-180 MYBP
UDC:551.7+"621.2"	Cambrian. 600-490 MYBP
UDC:551.7+"621.3"	Ordovician. 490-430 MYBP
UDC:551.7+"621.4"	Silurian. Gothlandian. 430-400 MYBP
UDC:551.7+"621.5"	Devonian. 400-350 MYBP
UDC:551.7+"621.6"	Carboniferous. 350-270 MYBP
UDC:551.7+"621.7"	Permian. 270-220 MYBP
UDC:551.7+"622.2"	Triassic. 220-180 MYBP
UDC:551.7+"622.4"	Jurassic. 180-135 MYBP
UDC:551.7+"622.6"	Cretaceous. 135-70 MYBP
UDC:551.7+"628"	Cenozoic (Cainozoic). Neozoic
UDC:551.7+"628"	Tertiary. 70-1 MYBP
UDC:551.7+"628.2"	Palaeogenic. Nummulitic
UDC:551.7+"628.22"	Palaeocene
UDC:551.7+"628.24"	Eocene
UDC:551.7+"628.26"	Oligocene
UDC:551.7+"628.4"	Neogene
UDC:551.7+"628.42"	Miocene
UDC:551.7+"628.44"	Pliocene
UDC:551.7+"628.6"	Quaternary. 1 MYBP - Present
UDC:551.7+"628.62"	Pleistocene in general. Diluvium
UDC:551.7+"628.64"	Holocene. Postglacial in general

2) Space: Table VI shows an excerpt of the resulting UDC classification practically used for spatial features and place.

TABLE VI. UNIVERSAL DECIMAL CLASSIFICATION OF SPATIAL FEATURES AND PLACE USED WITH THE KNOWLEDGE RESOURCES (EXCERPT).

UDC Code	*Description (English)*
UDC:(1)	Place and space in general. Localization. Orientation
UDC:(1-0/-9)	Special auxiliary subdivision for boundaries and spatial ...
UDC:(1-0)	Zones
UDC:(1-1)	Orientation. Points of the compass. Relative position
UDC:(1-19)	Relative location, direction and orientation
UDC:(1-2)	Lowest administrative units. Localities
UDC:(1-5)	Dependent or semi-dependent territories
UDC:(1-6)	States or groupings of states from various points of view
UDC:(1-7)	Places and areas according to privacy, publicness ...
UDC:(1-8)	Location. Source. Transit. Destination
UDC:(1-9)	Regionalization according to specialized points of view
UDC:(2)	Physiographic designation
UDC:(20)	Ecosphere
UDC:(21)	Surface of the Earth in general. Land areas in particular. ...
UDC:(23)	Above sea level. Surface relief. Above ground generally. ...
UDC:(24)	Below sea level. Underground. Subterranean
UDC:(25)	Natural flat ground (at, above or below sea level). ...
UDC:(26)	Oceans, seas and interconnections
UDC:(28)	Inland waters
UDC:(29)	The world according to physiographic features
UDC:(3)	Places of the ancient and mediaeval world
UDC:(32)	Ancient Egypt
UDC:(36)	Regions of the so-called barbarians
UDC:(37)	Italia. Ancient Rome and Italy
UDC:(38)	Ancient Greece
UDC:(4/9)	Countries and places of the modern world

Classification references of that kind have been implemented with knowledge resources and geo-coordinates.

Any of the classification can be mapped to specific content data. The workflows and processing handle different dates and specification between classification and content as well as using equal classification elements for different absolute dates, e.g., as required for different regions or cultures.

3) Creation of knowledge objects: The following examples are taken from astrophysical, volcanological, and geoscientific context, all referring to a large number of natural sciences' disciplines and humanities. Especially, types of meteorites and types of volcanoes do have different origin but the context links them in arbitrary ways in space and time. Examples are their linkage in geological time scale, location, geological attributes as well as their association with cultural events and secondary use of their material in many objects or buildings. An object carousel generated for impact craters, shows the different types present in the knowledge resources groups and their crater categories (Figure 1).

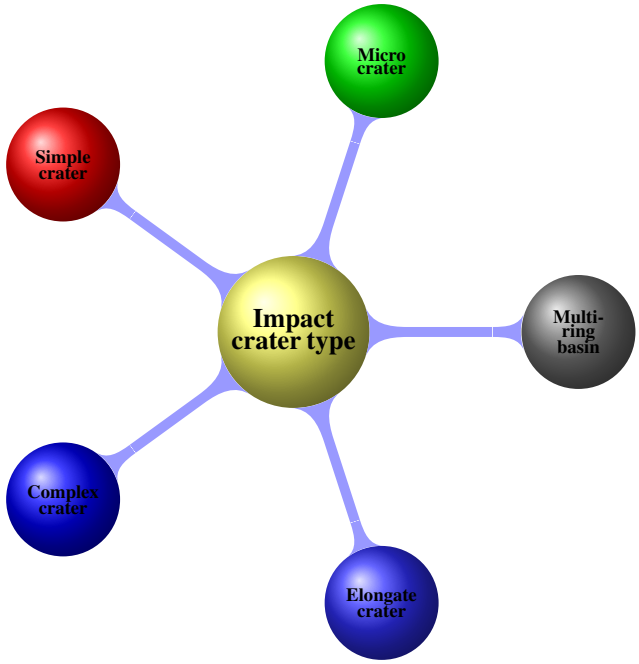


Figure 1. Object Carousel computed for impact crater categories.

Any objects in the categories can carry attributes like time and space as well as objects in other categories, which allows to have dimensions across disciplines like in the following example with impact craters and volcanoes.

4) Results of systematical use: Suitable views for volcanic features are: Type (of volcano, coarse categories), date on timeline, size (height). For craters respective views are: Type (of crater, fragmentary), date on timeline, size (diameter). Two Object Carousels have been computed. Figure 2 shows the knowledge resources groups for volcano types, and Figure 3 provides the geological spaces of time references.

Figure 2. Object Carousel for volcano and type references computed for terrestrial volcanism, providing volcano type references.

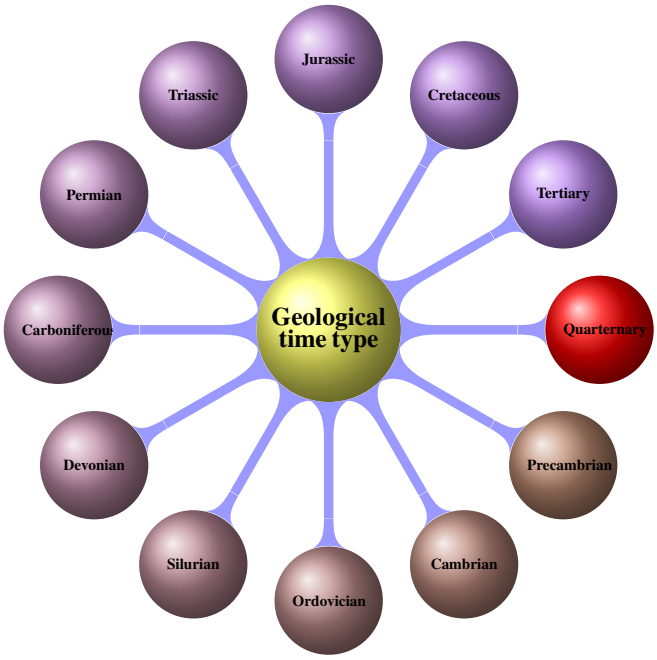


Figure 3. Object Carousel on geological spaces of time for computed references (terrestrial volcanoes, impact craters, and geological processes).

For simplicity only the main groups are shown, subgroups like for Quarternary "Holocene" and "Pleistocene" create separate carousels (Figure 4). Most geological objects have references into some instance of these carousels. This enables to create numberless links to additional information.

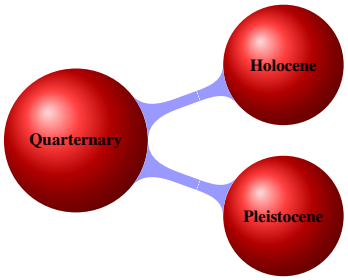


Figure 4. Object Carousel "Quarternary".

The colour coding for Carousels is symbolic and can be defined to represent any grouping as decided within the workflow. It can result from the grade of detail required for the description. In this case, the colour red links the three shown Object Carousels with the information referring to a requested object like "Vesuvius". The subgroup Object Carousels, e.g., "Quarternary" (Figure 4), opens additional references to volcanological feature objects. The listing in Figure 5 shows context replacement definitions and corrections.

```
1  Cretacious :: Cretaceous
2  Kreide :: Cretaceous
3  Trias :: Triassic
4  Carbon :: Carboniferous
5  Karbon :: Carboniferous
6  Silurium :: Silurian
7  Silur :: Silurian
8  Ordovicium :: Ordovician
9  Ordovizium :: Ordovician
10 Cambrium :: Cambrian
11 Kambrium :: Cambrian
12 Precambrium :: Precambrian
13 Präkambrium :: Precambrian
```

Figure 5. Replacement definition for relevant terms (LX resources).

The example lists an excerpt of relevant terms and types of notation that can be considered equal for the target context.

5) Processing media citation references: Figure 6 shows an excerpt of a media citation set used with UDC classified knowledge objects, here with a Vesuvius reference.

```
1 cite: YES 20070000 {LXK:Pompeii; Vesuvius; reconstruction
  ; 3D; animation; Holocene} {UDC:...} {PAGE:----..----}
  LXCITE://Bonaventura:2007:My_DVD
2 cite: YES 20130000 {LXK:Pompeii; Vesuvius; Vesuvio;
  Holocene; postcard} {UDC:...} {PAGE:----..----} LXCITE:
  //Guardasole:2013:Vesuvio_1270m
3 cite: YES 20070000 {LXK:Pompeii; Vesuvius; reconstruction
  ; diorama} {UDC:...} {PAGE:----..----} LXCITE://
  Bonaventura:2007:Pompeii
4 cite: YES 20070000 {LXK:Pompeii; Vesuvius; bakery; mill
  stones; material; stone; volcanic lava; basalt; Holocene
  ; diorama} {UDC:...} {PAGE:--56..--59} LXCITE://
  Bonaventura:2007:Pompeii
```

Figure 6. Media citation set excerpt used with the UDC classified knowledge object "Vesuvius" (LX resources).

The examples are part of the "Vesuvius" and "volcanic mill stone" object references. The media citations refer to 3D video animations and dioramic reconstructions as well as even to postcards. These references resolve to [49] (animation), [50] (postcard), [51] (diorama).

Objects can carry an arbitrary number of classifications views, e.g., from automated classification to individual classification by different groups of experts. The facets itself are to be built from a base of universal classification, which is under continuous development and fully consistent in its editions.

For example, knowledge can be created by a single researcher, a research institute, a collaborative effort or any other process. Knowledge resources can therefore be created by a single group or they may be created by larger organisations. Taking library and museum scenarios as examples, then the practice is to have editions representing classification states as well as instances of objects. The resources and objects can use any number of these editions and instances.

6) Classification development: All classifications are subject of a continuous development, review, and auditing process. This is also true for the UDC itself, independent from its use for different disciplines or scenarios. Table VII shows an example in different UDC editions.

TABLE VII. DEVELOPMENT OF "TERTIARY" CLASSIFICATION WITH UDC EDITIONS AND KNOWLEDGE RESOURCES (EXCERPT).

UDC Code (a)	*UDC Code (b)*	*Description*
UDC:"623"	UDC:"628"	Tertiary (70-1 MYBP)
UDC:"623.1"	UDC:"628.2"	Palaeogene (70-25 MYBP)
UDC:"623.5"	UDC:"628.4"	Neocene (25-1 MYBP)
UDC:551.77	UDC:551.7+"628"	Cenozoic (Cainozoic). Neozoic
UDC:551.78	UDC:551.7+"628"	Tertiary. 70-1 MYBP
UDC:551.781	UDC:551.7+"628.2"	Palaeogenic. Nummulitic
UDC:551.781.3	UDC:551.7+"628.22"	Palaeocene
UDC:551.781.4	UDC:551.7+"628.24"	Eocene
UDC:551.781.5	UDC:551.7+"628.26"	Oligocene
UDC:551.782	UDC:551.7+"628.4"	Neogene
UDC:551.782.1	UDC:551.7+"628.42"	Miocene
UDC:551.782.2	UDC:551.7+"628.44"	Pliocene

The example is the "Tertiary" classification development within different UDC editions. The table shows that the target not only moved $(a) \rightarrow (b)$ within the classification but was also adapted to a new subgrouping (lower block). The currently final result is a composite classification, composing from geology and time, holding both Tertiary and Cenozoic.

UDC still not considers different stratigraphies in plain. Further, respective editions can be used or references to respective editions and entries. The editions are fully consistent in themselves, so it is natural that the overall consistency of workflows using different editions has to be cared for by the disciplines or providers.

Figure 7 shows Object Carousels computed for a complete common system (top) as well as for an alternative system (below) used for some purposes [43] after the year 2000, missing "Tertiary". The colours represent the term levels within the respective system.

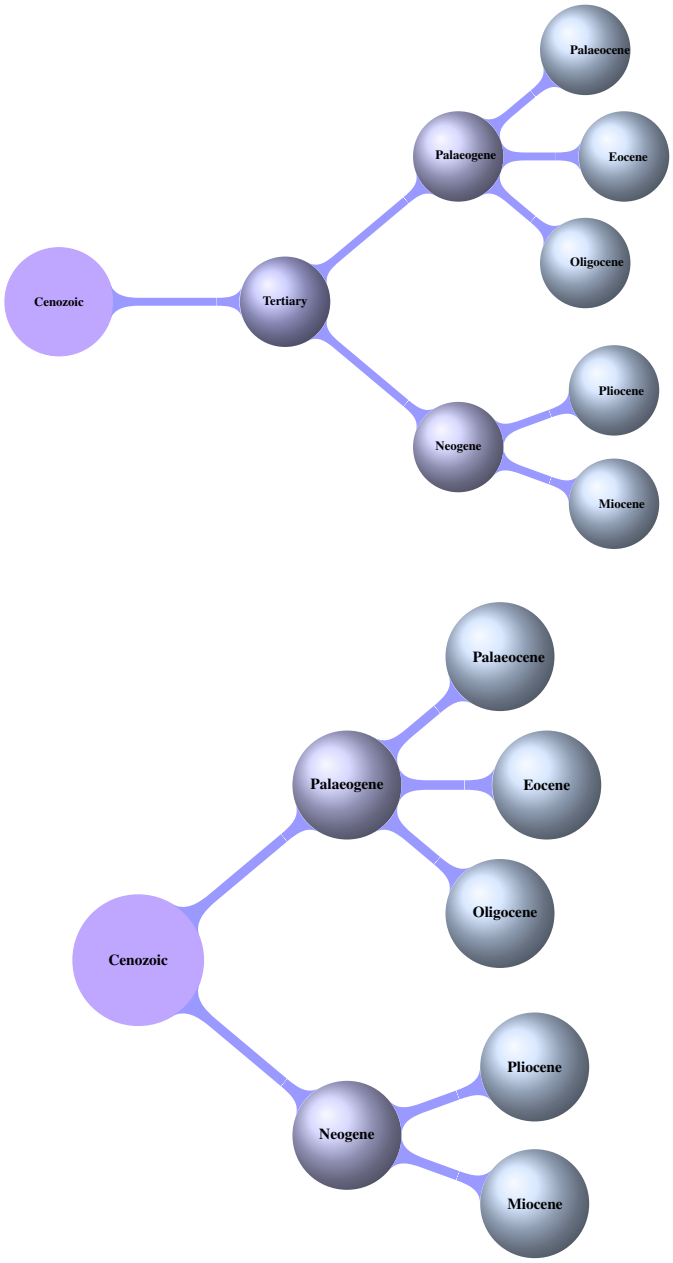


Figure 7. Object Carousel "Tertiary": Common (top) and alternative (below).

Moved items have to be considered "persistent" within long-term knowledge resources appropriately with all consequences. It is possible to support any number of instances or versions within the knowledge resources as long as each is handled consistently. In this case, the consistency can refer to the classification, the content, as well as to the references. The UDC can reflect the developments in classification with editions, which can be used consistently with the resources.

7) Result matrix: Table VIII shows the results from the computation of a systematical classification of volcanological features, short "volcano types".

TABLE VIII. COMPUTED SYSTEMATICAL CLASSIFICATION OF VOLCANOLOGICAL FEATURES FROM THE KNOWLEDGE RESOURCES.

Volcano Type	*Group*	*References Data Examples*
Complex volcano	A	Vesuvius VNUM:0101-02= UDC:[551.21+911.2+55]:[902]"63"(4+23+24)... GPS:40.821N14.426E Quarternary VEI:VEI5
Compound volcano	A	Cayambe VNUM:1502-004 UDC:[551.21+911.2+55]:(8+23+24)... GPS:... Holocene ...
Somma volcano	A	Ebeko VNUM:0900-38= UDC:... GPS:... Quarternary ...
Submarine volcano	A	Campi Flegrei Mar Sicilia VNUM:0101-07= UDC:... GPS:... Quarternary ...
Subglacial volcano	A	Katla VNUM:1702-03= UDC:... GPS:... Quarternary ...
Unspecified type	A	– VNUM:– GPS:... – ...
Strato volcano	B	Vulcano VNUM:0101-05= UDC:... GPS:... Quarternary ...
Shield volcano	C	Etna VNUM:0101-06= UDC:... GPS:... Quarternary ...
Explosion crater	D	Larderello VNUM:0101-001 UDC:... GPS:... Quarternary ...
Caldera	D	Campi Flegrei VNUM:0101-01= UDC:... GPS:... Quarternary ...
Tuff cone	E	Tutuila VNUM:0404-02- UDC:... GPS:... Holocene ...
Scoria cone	E	Antofagasta de la Sierra VNUM:1505-124 UDC:... GPS:... Holocene ...
Pyroclastic cone	E	Anunciacion, Cerro VNUM:1405-032 UDC:... GPS:... Holocene ...
Cinder cone	E	Chiquimula Field VNUM:1402-20- UDC:... GPS:... Holocene ...
Lava dome	E	El Chichon VNUM:1401-12 UDC:... GPS:... Quarternary ...
Volcanic field	F	Holotepec VNUM:1401-07- UDC:... GPS:... Quarternary ...
Hydrothermal field	F	Musa River VNUM:0503-02= UDC:... GPS:... Quarternary ...
Fumarole field	F	Kos VNUM:0102-06= UDC:... GPS:... Pleistocene ...
Maar	F	West Eifel Volcanic Field VNUM:0100-01- UDC:... GPS:... Quarternary ...
Fissure vent	F	Quetena VNUM:1505-074 UDC:... GPS:... Holocene ...

It compiles a small excerpt of computed data from the LX resources [33]. The table delivers comprehensive information for the volcanological topics integrated here: Volcanic feature types, computed groups, UDC mappings, and examples of computed references, e.g., Volcano Number (VNUM) the volcanic reference file number, geo-coordinates and spatial data, and spaces of time, as well as referenced data, e.g., the Volcanic Explosivity Index (VEI) [52]. The full result matrix for this request contains several hundreds of computed objects with tenthousands of references.

The selection of the most relevant objects is not an issue of the documentation or the view itself. The selection must be handled by the algorithms and workflows from the disciplines handling the specific resources, e.g., for providing a ranking. Disciplines can be any, therefore the algorithms can comprise statistics as well as ranking algorithms or image processing techniques, depending on the objects and their features.

In this case the most relevant elements are defined by the objects in the volcanological feature container. In addition, in-depth completion within object containers has been enabled for the case of volcanological features. A container represents a collection of equally structured groups of related objects on a certain topic. Examples for available containers are:

- Volcanological object container,
- Earthquake object container,
- Meteoric impact object container,
- Astronomy object container,
- Mineral object container.

The in-depth completion allows additional data, e.g., spatial data, processing, and media object references.

The resources further allow for a flexible mapping of attributes, e.g., container relations, classification, keywords, numbers, references, media samples, material samples, spatial data, and geological spaces of time. With these references the volcanological features can be associated with a VEI, e.g., Vesuvius (Pompeii) VEI5, Krakatau VEI6, Tambora VEI7, Thera (Santorini) VEI7, Toba (Sumatra) VEI8, whereas a "Caldera" object itself being a crater does not have a VEI.

With existing models used in simulation and modelling there is no consideration of references between disciplines, e.g., volcanoes and weather. With the knowledge resources, volcanological features can be referred to volcanological events, seismological events, and weather phenomenon events or biology. The larger the data base is the more correlatable events get available in space and time. In comparison to mono-disciplinary information the multi-disciplinary context of the knowledge resources supports an improved knowledge description. Further, even indirect correlation, e.g., in the above case between volcanic features and meteorite impact features can be investigated.

8) Knowledge generation, combination, and visualisation: The following visualisation (Figure 8) paradigmatically illustrates the results from the compute requests. An on-location attribute has been chosen for the relations in order to compute a distribution map for volcanic features using the `lxlocation` workflow. The location attribute is suitable for referring to an unlimited number of multi-disciplinary information in this case.

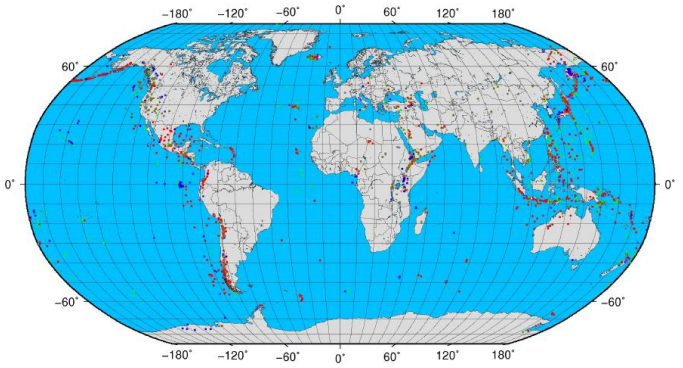


Figure 8. Volcanomap – computed worldwide spatial distribution of classified terrestrial volcanological features from resulting object entries.

The distribution is computed from the result matrix of related object context of several hundred classified terrestrial volcanic features via the knowledge resources research database. The result matrix is the result of the present content, references, and workflow. In all examples only an excerpt of these can be shown. Several modules have been used for this example: `select_knowledge_environment`, `lxgrep_udc`, `lxkwgrep`, `generateCarousel`, `lxvolcanoes2gmt`, `cprgmt_world_cprvolcanoes_separated`, as well as `pscoast`, `pstoraster`, and `psxy`. System interfaces can be created via instructions, programming interfaces, or any kind of interface the disciplines working on implementations and suggested workflows want to built on top of the knowledge resources. The workflow allows any feature supported by the deployed components, e.g.,

- Association by classification weighting,
- Association by grouping,
- Association by colourisation,
- Association and by symbolisation.

Any association and context that can be described and expressed for any objects can be realised with the knowledge resources. In this example, colour groups have been computed via the result matrix (Table VIII): A: green, B: red, C: blue, D: lighter blue, E: grey, F: dark green. The volcanic features are classified and several classification groups have been chosen for the result. The map shows the present situation according to the present state of the available volcanological data.

If we want to generate an according section view on data from the geo-related knowledge resources, we can choose from a number of topics and disciplines. A comparable section view regarding the dimensions and attributes can be described on content base (object classification) by a set of

- Planetary surface objects,
- Geological / geophysical context,
- Geo-coordinates,
- Spatial distribution.

The creator of the overall workflow can define what to do with the data and how to present it. This includes adding a suitable application component for representing the knowledge for the the intended purpose. From application component base, in this case GMT, it can be described by

- Mapping,
- Projection,
- Grid parameters,
- Label attributes / size,
- Colours.

The component specific description will most probably be placed with the workflow implementation but if required it might also be integrated with the objects. The result is shown in the associated sample distribution of terrestrial impact features (meteorite) as depicted in Figure 9.

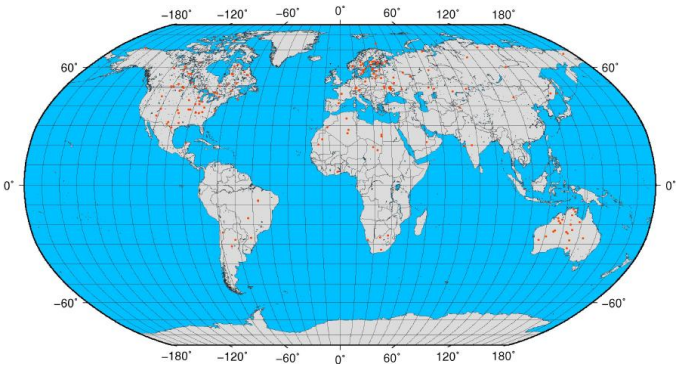


Figure 9. Impactmap – computed worldwide spatial distribution of classified terrestrial impact features (meteorite) from resulting object entries.

The map is computed from the related object context (`lxlocation` workflow) contained in a meteorite impacts features research database of the knowledge resources. It is possible to combine any information, e.g., computing a map animation varying in time, showing the development of volcanic features. Therefore, a temporal sequence of spatial distributions can be used as a simple example for variations over time.

On base of a physical view (criteria classification) the criteria for impact crater classification are:

- Size of the impacting object,
- Speed of the impacting object,
- Material of the impacting object,
- Composition and structure of the target rock,
- Angle that the impacting object hits the target,
- Gravity of the target object respective planet,
- Porosity and other ... of the impacting object,
- Age of the impact,
- Size of the impact,
- Structure of the crater.

Further associated phenomena (indicator classification) are impact crater indicators on the other hand, which are:

- Planar fractures in quartz,
- Shocked quartz,
- Glass fragments.

When approaching from the “catastrophe” view it has shown that the most prominent relation is the “size”. This mostly correlates with “diameter” and still mapping and timelining will come natural.

9) Processing and computational numbers and issues: Table IX shows the processing and computational demands per instance resulting from the presented scenarios.

TABLE IX. PROCESSING AND COMPUTATIONAL DEMANDS.

Item	*Value / Description*
UDC, number of classification items	70,000
Number of classification languages	50
Number of classification variations (50×70,000)	3,500,000
Knowledge object subset, number of items	100,000
Number of terms	10,000,000
Number of object languages	2
Operations, number per subset result entry	50,000,000
Number per subset result entry, incl. keywords	500,000,000
Parallelisation (subset), wall time / num. of nodes	7,500 s / 1
Wall time / number of nodes	1,300 s / 10
Wall time / number of nodes	220 s / 100
Wall time / number of nodes (extrapolated)	4 s / 10,000

Besides the large requirements per instance with most workflows there are significant effects by parallelising even within single instances. The following issues have shown to lead to advanced challenges and increased processing and computational times. Nomenclature, terms, and attributes tend to be at least partially different in different cultures and languages. For many discovery workflows as well as efforts to increase the quality of the result matrices it is necessary to consider more than one culture and language. Processing a classification numbering in decreasing numbering with increasing age or following in different directions is less consequent. For example, in geosciences it is natural to start spaces of time with Quarternary, followed by older stratigraphy. In addition to the existing singular spaces of time mapping most objects require appropriate different mappings to absolute dates, e.g., with Bronze Age having different absolute dates for different regions or cultures. The calculation with extensive composite classifications, facets, and respective ranges instead of native classifications can increase the computational requirements drastically as has been shown with the knowledge content from the Gottfried Wilhelm Leibniz resources [36].

VIII. KNOWLEDGE INTEGRATION: CONCEPTUAL CONTRIBUTION

Integrating referenced knowledge resources can essentially contribute valuable content, both on quantity and depth. Knowledge dimensions as well as computational views can benefit from these contributions.

Nevertheless, references to unstructured or differently structured knowledge objects will contain a number of deviations, which may need to be unified, either for contributing to manifoldness or to intensification of resulting statements.

Utilising the classifications and facets for collecting references to material can vastly enrich the matrices and results, which has already been proposed [53] for solutions coping with Big Data resources [54]. This procedure can help to overcome conceptual deficits of unstructured material as well as language aspects. The procedure delivers a high percentage of relevant otherwise unstructured conceptual knowledge and allows for an efficient unification and integration of the results.

One of the largest resource of references to unstructured, heterogeneous, multi-lingual data are services like Google [55]. Table X shows the results resolved from references of a Google search done for the topics “volcano, udc, classification”. This means references for classifications and other conceptual knowledge can be used, in any combination, in order to trigger search requests. The results contain the UDC classification found with the request as well as the terms associated with this in the text.

TABLE X. VOLCANO RESULTS FROM PUBLICLY AVAILABLE INFORMATION, GOOGLE SEARCH, STATUS OF JANUARY 2013 (EXCERPT).

UDC Classification	*In-text Terms*
UDC:551.442(437.6)	Volcano phreatic
UDC:631.4	Volcano
UDC:553.405	Uranium deposit volcano
UDC:551.31:551.44(532)	Volcano
UDC:(*764)	Volcano
UDC:(*7)	Volcano

Table XI shows the results resolved from references of a Google search done for the topics “cenote, udc, classification”.

TABLE XI. CENOTE RESULTS FROM PUBLICLY AVAILABLE INFORMATION, GOOGLE SEARCH, STATUS OF JANUARY 2013 (EXCERPT).

UDC Classification	*In-text Terms*
UDC:930.85(726.6) 551.435.8:528.9	Sinkhole cenote maya
UDC:551.44	Doline sinkhole
UDC:556.34:519.216	Sinkhole drainage
UDC:551.435.82(234.41)	Sinkhole collapse
UDC:624.153.6:699.8:551.448	Sinkhole collapse
UDC:551.44(450.75)	Karst sinkhole collapse
UDC:551.44(045)=20	Groundwater surfacewater
UDC:551.44:001.4	Grotte Höhle
UDC:551.44(450.75)	Karst Apulia
UDC:551.44(437.2)	Geology karst

All documents found from public external sources have been identified to contain academical and scientific content. Even as this example is intended to provide a lower depth of knowledge mapping than available in specialised knowledge resources, it provides an excellent spectrum of related information for the respective disciplines. First, the unstructured “Big Data” Google is using is currently not classificed and, second, Google cannot search using a universal classification up to now. This means that up to now most conceptual information request cannot be expressed directly via search engines.

Digital sources can be consistently classified in groups as well as single objects. Groups or objects and their associated references can refer to a classification. The examples show the possible bandwidth within a topic. They show that there is much more in-depth knowledge in the data. The classification support can extract focussed knowledge even if the data is only available in different languages as in this case. Therefore, internal as well as external sources have been used for the examples.

IX. Discussion and evaluation

The results presented here have several scientific and technical aspects. The following passages discuss some major contributions to the content and scientific results as well as to the knowledge resources and features.

A. Knowledge resources

The Knowledge Oriented Architecture (KOA) of the resources is based on a flexible integration of the documentation and development architecture utilising the Collaboration house framework for disciplines, services, and resources [34]. The knowledge objects, here the geological and volcanic feature objects, can be used with any of their attributes. Therefore, any references to objects belonging or referring to any other objects can be computed from this. For an object referred to a timescale of periods other objects can be associated with the respective object, even beyond direct references. For example, "geological time type" can refer from "volcano type" to any other suitable for a geological or comparable spaces of time classification. For example, this will be true for geophysical, palaeontological or archaeological objects. Further, volcanic objects from the Quarternary can be associated to meteorite impact events from the Quarternary. The more, they can be restricted to associated objects of a certain attribute, e.g., from the same region. With secondary steps further information can be integrated. This can include geophysical data, media data or associated objects. The resulting quality depends on an intelligent use of context and classification. A strong classification support is essential, the more as object and even many citations, media, and publications are not explicitly aware of the nomenclature of spaces of time used with specific content can, e.g., to express that the spaces of time refer to plants or animals. Employing a universal classification with multidisciplinary content this way, e.g., with volcanological content, expedites knowledge discovery as well as it targets on scientific discovery.

Regarding methodology it further allows to

- Support a systematic documentation,
- Define a normative classification,
- Define cognitive interfaces.

Regarding architecture and implementation it allows to

- Support decision making in complex systems,
- Implement learning system components and
- Support components by intelligent systems.

Creating classified knowledge resources objects has proven to be most sustainable for a significant period of operation and implementation. It has been efficient and portable with all application scenarios and environments for more than two decades, used with ten different operating system environments, with different editing components, processing languages, and compute and storage resources. From classification side it is suggested to have advanced computing support, e.g., for spaces of time as well as for the complementary systematics for disciplines. In addition, a methodological framework for UDC supporting the required processing and computation would add immense benefits to its universal applicability. Some new types of stratigraphies have not widely been adopted and should again become subject of modification regarding a long-term use. In many cases, the consequence of claims on consistency has been to use one dedicated edition of the classification. This shall ensure consistency within the application. Using a small subset of classification can help to reduce the appearent work that has to be done for classification but it cannot ensure to avoid variances in different editions. Consistent version management support for the classification has shown to become necessary as soon as knowledge resources are using modified classifications over time.

B. Content: Case lessons learned

Besides the detailed results and references, the overall results for the discussed cases are:

- Volcanic features are well known above and below sea-level and are more often long-term processes.
- Known impact features show a concentration in highly populated and industrialised areas.
- Impact features have been reduced by morphological processes and are mostly only available above sea-level.
- Both impact and volcanic features are related to social and archaeological findings.
- Both impact and volcanic features are publicly known.
- Compared with impacts and volcanic features, archaeological sites and results are not known to the same amount in order to protect the sites.

C. Classification expenses and benefits

Classification support, e.g., via UDC, does require intensive work and can be expensive if used in non automated ways. Nevertheless, this can make a difference as classification views very much profit from professional experiences. The application of UDC with complex knowledge context requires flexibility of the resources as wells as a flexible handling regarding extendability. The challenges with the distributed use of UDC are, e.g., the use of private catalogues, like external codes or author abbreviations. In addition, the sustainability of knowledge objects benefits from the use of methods like facetted versioning, universal dates (e.g., ISO dates), and georeferencing.

D. Complementary information and classification

Text information and classification information are complementary. This is important for knowledge resources as well a for application scenarios, e.g., search algorithms. Using classification supported search algorithms can improve the result drastically. The quality of results improves from below five percent to up to over ninety percent.

With the presented Object Carousels an undefined number of practical workflows can be created on the knowledge resources. Examples, which have been investigated for gathering complementary results are regular expression and string search, classification search (UDC), keyword search, sort support search, references search or phonetic search (Soundex).

E. Knowledge and views

UDC can be used with any object in any context and can help to reduce compute requirements with knowledge discovery.

The application of a universal classification and knowledge resources can drastically reduce the computational requirements, as well as it supports the parallelisation of instances within workflows due to the large numbers of representations in common per-instance calls (Table IV). The classification views and facets enable to reduce the amount of object analysis required for discovery and reuse workflows. The algorithms benefitting from this are on the algorithm side (for example, object and references search, content string comparisons, and knowledge based associations) on the hardware and resources side (for example, input/output requests including read/write processes, compute resources' and memory requirements, and communication requirements).

X. Conclusion and future work

It has been demonstrated that multi-disciplinary knowledge-based objects can be successfully used in order to create computational views. The knowledge processing employing UDC classification has shown to be a universal and most flexible solution for creating long-term multi-disciplinary knowledge resources and providing a base for universal knowledge dimensions.

The implemented knowledge resources as well as deployed conceptual knowledge, demands on processing and resources, and examples for the computational views have been presented, delivering components that can be used with future developments. In addition, the examples present content and references on about more than 1,000 objects (volcanological features and terrestrial impact features), which delivered a base for new discovery after having been integrated with the knowledge resources.

The resources and framework can be used even with basic attributes and cross-references, and assure support for subsequent use and knowledge procurement processes. Structuring and classification with long-term knowledge resources and UDC support have successfully provided excellent solutions, which can be used for natural sciences, e.g., geosciences, volcanology or with spatial disciplines as well as for universal knowledge. The knowledge resources can provide any kind of Object Carousel and object references. Decisions can be computed with support of the UDC classification. Due to the universal long-term multi-disciplinary knowledge gathering, the knowledge resources are a general universal decision support base.

Besides these, a major benefit of the extensive support of UDC language translations is that regarding discovery workflows it can also be used for improving the quality as well as the quantity of elements and references in the result matrices. Employing a universal classification when creating knowledge resources has provided substantial benefit for both. The workflow procedures build for special purposes are property of the researchers and disciplines creating, developing, and operating their implementations. The data used by them is intended to be part of the respective collaboration. Currently, if someone creates data, he or she can use the data and share it with others, creating agreements and policies.

As knowledge resources have been proven to be a valuable means for research in many disciplines, components are candidates for community tasks as well as for open access development and licensing models. Currently, the policies with many collaborations, funding, and services (as comparable with the UDC model) do not allow to make sources and content of the knowledge resources public.

Because the process of creating long-term sustainable content is quite pretentious and will never be completed there might be support by a sustainable funding in the future, too. No discipline should commonly be funded because it is possible to increase the disciplines' requirements for associated tasks. Generally, the priority should be on knowledge "content and result" and not merely on the quantity of background resources required.

On the other end, the operation for disciplines, services, and resources providers can be accompanied by licensing models, supporting a sustainable long-term development and operation on all three sections. Factual and conceptual knowledge, e.g., information collections and classification editions, web services and interfaces, e.g., discovery and section views, as well as compute and storage services can be provided, developed, and priced that way.

As presented, the knowledge processing can base on a solid and sustainable long-term resource, which allows to create any kind of workflows, dynamical discovery, and IICS components and facilitate the use of High End Computing resources. Based on this research, in the future further features for a high end technical integration of data interfaces and resources and more intelligent learning components can be developed.

Acknowledgements

We are grateful to all national and international partners in the GEXI cooperations for the innovative constructive work. We thank the Science and High Performance Supercomputing Centre (SHPSC) for long-term support of collaborative research since 1997, including the GEXI developments and case studies on archaeological and geoscientific information systems. Special thanks go to the scientific colleagues at the Gottfried Wilhelm Leibniz Bibliothek (GWLB) Hannover, especially Dr. Friedrich Hülsmann, for prolific discussion within the "Knowledge in Motion" project, for inspiration, and practical case studies. Many thanks go to the scientific colleagues at the Leibniz Universität Hannover, especially Mrs. Birgit Gersbeck-Schierholz, to the Institute for Legal Informatics (IRI), Leibniz Universität Hannover, and to the Westfälische Wilhelms-Universität (WWU), for discussion, support, and sharing experiences on collaborative computing and knowledge resources and for participating in fruitful case studies as well as to the participants of the postgraduate European Legal Informatics Study Programme (EULISP) for prolific discussion of scientific, legal, and technical aspects over the last years.

REFERENCES

[1] C.-P. Rückemann, "Knowledge Processing for Geosciences, Volcanology, and Spatial Sciences Employing Universal Classification," in Proceedings of The Sixth International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2014), March 23–27, 2014, Barcelona, Spain. XPS Press, 2014, pp. 76–82, ISSN: 2308-393X, ISBN: 978-1-61208-326-1, URL: http://www.thinkmind.org/download.php?articleid=geoprocessing_2014_4_10_30044 [accessed: 2014-11-30].

[2] A. Woodie, "Forget the Algorithms and Start Cleaning Your Data," Datanami, 2014, March 26, 2014, URL: http://www.datanami.com/datanami/2014-03-26/forget_the_algorithms_and_start_cleaning_your_data.html [accessed: 2014-11-30].

[3] "Universal Decimal Classification Consortium (UDCC)," 2014, URL: http://www.udcc.org [accessed: 2014-11-30].

[4] "Universal Decimal Classification (UDC)," 2014, Wikipedia, URL: http://en.wikipedia.org/wiki/Universal_Decimal_Classification [accessed: 2014-11-30].

[5] A. Slavic, "UDC libraries in the world - 2012 study," universaldecimalclassification.blogspot.de, 2012, Monday, 20 August 2012, URL: http://universaldecimalclassification.blogspot.de/2012/08/udc-libraries-in-world-2012-study.html [accessed: 2014-11-30].

[6] "Wissensmanagement in Bibliotheken," library essentials, LE_Informationsdienst, April 2014, 2014, pp. 9–11, ISSN: 2194-0126, URL: http://www.libess.de [accessed: 2014-11-30].

[7] S. Yadagiri and T. K. Gireesh Kumar, Knowledge Management: Changing Role of LIS in the Digital Environment. B. R. Publishing Corporation, Delhi, 2014, in: Libraries in the Changing Dimensions of Digital Technology, Commemorative Publication in Honour of Prof. D. Chandran, 2013, pp. 476–481, URL: http://eprints.rclis.org/22813 [accessed: 2014-11-30].

[8] C. Sherman, "What's the Big Deal About Big Data?" Online Searcher, 2014, pp. 10–16, March/April, 2014.

[9] "Was ist so „Big" an Big Data?" library essentials, LE_Informationsdienst, April 2014, 2014, pp. 6–8, ISSN: 2194-0126, URL: http://www.libess.de [accessed: 2014-11-30].

[10] Fundamentals of Library of Congress Classification, Developed by the ALCTS/CCS-PCC Task Force on Library of Congress Classification Training, 2007, Robare, L., Arakawa, S., Frank, P., and Trumble, B. (eds.), ISBN: 0-8444-1186-8 (Instructor Manual), ISBN: 0-8444-1191-4 (Trainee Manual), URL: http://www.loc.gov/catworkshop/courses/fundamentalslcc/pdf/classify-trnee-manual.pdf [accessed: 2014-10-30].

[11] O. Günther, "Big Data: Was ist das? Und was bedeutet es für Wissenschaft, Wirtschaft und Gesellschaft?" Informatik Spektrum, "Big Data", Springer Verlag, Berlin, Heidelberg, vol. 37, no. 2, Apr. 2014, pp. 85–86, DOI: 10.1007/s00287-014-0783-7.

[12] C. Leng, "Die Vorstandsperspektive: Big Data = Big Wealth?" Informatik Spektrum, "Big Data", Springer Verlag, Berlin, Heidelberg, vol. 37, no. 2, Apr. 2014, pp. 88–89, DOI: 10.1007/s00287-014-0780-x.

[13] J.-C. Freytag, "Grundlagen und Visionen großer Forschungsfragen im Bereich Big Data," Informatik Spektrum, "Big Data", Springer Verlag, Berlin, Heidelberg, vol. 37, no. 2, Apr. 2014, pp. 97–104, DOI: 10.1007/s00287-014-0771-y.

[14] C. Meinel, "Big Data in Forschung und Lehre am HPI," Informatik Spektrum, "Big Data", Springer Verlag, Berlin, Heidelberg, vol. 37, no. 2, Apr. 2014, pp. 92–96, DOI: 10.1007/s00287-014-0773-9.

[15] P. Liggesmeyer, J. Dörr, and J. Heidrich, "Big Data in Smart Ecosystems," Informatik Spektrum, "Big Data", Springer Verlag, Berlin, Heidelberg, vol. 37, no. 2, Apr. 2014, pp. 105–111, DOI: 10.1007/s00287-014-0772-x.

[16] IBM, "Wie man Erkenntnisse aus Big Data optimiert," 2014, URL: http://www.ibm.com/smarterplanet/de/de/madewithibm/stories/#!story/18?ref=home&cmp=333ab&ct=333ab02w&cr=google&cm=k&csr=41429analytics_and_big_data-astron&ccy=de&ck=big_data_storage&cs=broad&S_PKG=-&S_TACT=333ab02w&mkwid=sRUgxqgp6-dc_41984993492_432i044571 [accessed: 2014-11-30].

[17] S. Fischer, "Big Data: Herausforderungen und Potenziale für deutsche Softwareunternehmen," Informatik Spektrum, "Big Data", Springer Verlag, Berlin, Heidelberg, vol. 37, no. 2, Apr. 2014, pp. 112–119, DOI: 10.1007/s00287-014-0770-z.

[18] G. Dueck, "Bigger Data – Bigger Trouble?" Informatik Spektrum, "Big Data", Springer Verlag, Berlin, Heidelberg, vol. 37, no. 2, Apr. 2014, pp. 139–143, DOI: 10.1007/s00287-014-0774-8.

[19] M. E. Jennex, "A Proposed Method for Assessing Knowledge Loss Risk with Departing Personnel," vol. 44, no. 2, 2014.

[20] "Wissensverlust vermeiden beim Abgang von Wissensarbeitern," library essentials, LE_Informationsdienst, Juni/Juli 2014, 2014, pp. 9–11, ISSN: 2194-0126, URL: http://www.libess.de [accessed: 2014-11-30].

[21] "International Expert Panel on Quality and Risk Assessment in Telecommunications Services, July 22, 2014, The Tenth Advanced International Conference on Telecommunications (AICT 2014), The Tenth International Conference on Internet Monitoring and Protection (ICIMP 2014), July 20–24, 2014, Paris, France," 2014, URL: http://www.iaria.org/conferences2014/ProgramAICT14.html [accessed: 2014-11-23], URL: http://www.iaria.org/conferences2014/filesAICT14/AICT2014_EXPERT_PANEL.pdf [accessed: 2014-11-23].

[22] C.-P. Rückemann, "Knowledge Integration for Scientific Classification and Computation," in The Fourth Symposium on Advanced Computation and Information in Natural and Applied Sciences, Proceedings of The 12th International Conference of Numerical Analysis and Applied Mathematics (ICNAAM), September 22–28, 2014, Rhodes, Greece, Proceedings of the American Institute of Physics (AIP), AIP Conference Proceedings. AIP Press, 2014, ISSN: 0094-243X.

[23] C.-P. Rückemann and B. F. S. Gersbeck-Schierholz, "Object Security and Verification for Integrated Information and Computing Systems," in Proceedings of the Fifth International Conference on Digital Society (ICDS 2011), Proceedings of the International Conference on Technical and Legal Aspects of the e-Society (CYBERLAWS 2011), February 23–28, 2011, Gosier, Guadeloupe, France / DigitalWorld 2011. XPS, 2011, pp. 1–6, ISBN: 978-1-61208-003-1, URL: http://www.thinkmind.org/download.php?articleid=cyberlaws_2011_1_10_70008 [accessed: 2014-11-30].

[24] C.-P. Rückemann, Sustainable Knowledge and Resources Management for Environmental Information and Computation. MacMillan, 2014.

[25] C.-P. Rückemann, "Fostering Environmental Protection via Sciences and Society: Key Knowledge and Climate Change," May, 2014, Contribution to ISSC, 2014.

[26] World Social Science Report 2013, Changing Global Environments, 1st ed. Published jointly by the United Nations Educational, Scientific and Cultural Organization (UNESCO), the International Social Science Council (ISSC), the Organisation for Economic Co-operation and Development (OECD), 2013, DOI: 10.1787/9789264203419-en, OECD ISBN 978-92-64-20340-2 (print), OECD ISBN 978-92-64-20341-9 (PDF), UNESCO ISBN 978-92-3-104254-6 (PDF and print).

[27] "The Digital Archaeological Record (tDAR)," 2014, URL: http://www.tdar.org [accessed: 2014-11-30].

[28] "WDL, World Digital Library," 2014, URL: http://www.wdl.org [accessed: 2014-11-30].

[29] Wissenschaftsrat, "Übergreifende Empfehlungen zu Informationsinfrastrukturen, (English: Spanning Recommendations for Information Infrastructures)," Wissenschaftsrat, Deutschland, (English: Science Council, Germany), Drs. 10466-11, Berlin, 28.01.2011, 2011, URL: http://www.wissenschaftsrat.de/download/archiv/10466-11.pdf [accessed: 2014-11-30].

[30] C.-P. Rückemann, "Implementation of Integrated Systems and Re-

sources for Information and Computing," in Proceedings of the International Conference on Advanced Communications and Computation (INFOCOMP 2011), October 23–29, 2011, Barcelona, Spain, 2011, pp. 1–7, ISBN: 978-1-61208-009-3, URL: http://www.thinkmind.org/download.php?articleid=infocomp_2011_1_10_10002 [accessed: 2014-11-30].

[31] C.-P. Rückemann, Queueing Aspects of Integrated Information and Computing Systems in Geosciences and Natural Sciences. In-Tech, 2011, pp. 1–26, Chapter 1, ISBN-13: 978-953-307-737-6, DOI: 10.5772/29337.

[32] "Geo Exploration and Information (GEXI)," 1996, 1999, 2010, 2014, URL: http://www.user.uni-hannover.de/cpr/x/rprojs/en/index.html#GEXI [accessed: 2014-10-26].

[33] "LX-Project," 2014, URL: http://www.user.uni-hannover.de/cpr/x/rprojs/en/#LX (Information) [accessed: 2014-10-26].

[34] C.-P. Rückemann, "Enabling Dynamical Use of Integrated Systems and Scientific Supercomputing Resources for Archaeological Information Systems," in Proceedings of The International Conference on Advanced Communications and Computation (INFOCOMP 2012), October 21–26, 2012, Venice, Italy. XPS, 2012, pp. 36–41, ISBN: 978-1-61208-226-4, URL: http://www.thinkmind.org/download.php?articleid=infocomp_2012_3_10_10012 [accessed: 2014-11-30].

[35] "UDC Online," 2014, URL: http://www.udc-hub.com/ [accessed: 2014-11-30].

[36] C.-P. Rückemann, "Archaeological and Geoscientific Objects used with Integrated Systems and Scientific Supercomputing Resources," International Journal on Advances in Systems and Measurements, vol. 6, no. 1&2, 2013, pp. 200–213, ISSN: 1942-261x, LCCN: 2008212470 (Library of Congress), URL: http://www.thinkmind.org/download.php?articleid=sysmea_v6_n12_2013_15 [accessed: 2014-11-23], URL: http://lccn.loc.gov/2008212470 [accessed: 2014-11-23].

[37] C.-P. Rückemann, "Sustainable Knowledge Resources Supporting Scientific Supercomputing for Archaeological and Geoscientific Information Systems," in Proceedings of The Third International Conference on Advanced Communications and Computation (INFOCOMP 2013), November 17–22, 2013, Lisbon, Portugal. XPS Press, 2013, pp. 55–60, ISSN: 2308-3484, ISBN: 978-1-61208-310-0, URL: http://www.thinkmind.org/download.php?articleid=infocomp_2012_3_10_10012 [accessed: 2014-11-30].

[38] C.-P. Rückemann, "High End Computing for Diffraction Amplitudes," in The Third Symposium on Advanced Computation and Information in Natural and Applied Sciences, Proceedings of The 11th International Conference of Numerical Analysis and Applied Mathematics (ICNAAM), September 21–27, 2013, Rhodes, Greece, Proceedings of the American Institute of Physics (AIP), AIP Conference Proceedings, vol. 1558. AIP Press, 2013, pp. 305–308, ISBN: 978-0-7354-1184-5, ISSN: 0094-243X, DOI: 10.1063/1.4825483.

[39] C.-P. Rückemann, "Advanced Scientific Computing and Multi-Disciplinary Documentation for Geosciences and Archaeology Information," in Proc. of The Int. Conf. on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2013), February 24 – March 1, 2013, Nice, France. XPS Press, 2013, pp. 81–88, ISSN: 2308-393X, ISBN: 978-1-61208-251-6, URL: http://www.thinkmind.org/download.php?articleid=geoprocessing_2013_4_10_30035 [accessed: 2014-11-30].

[40] U. Inden, D. T. Meridou, M.-E. C. Papadopoulou, A.-C. G. Anadiotis, and C.-P. Rückemann, "Complex Landscapes of Risk in Operations Systems Aspects of Processing and Modelling," in Proceedings of The Third International Conference on Advanced Communications and Computation (INFOCOMP 2013), November 17–22, 2013, Lisbon, Portugal. XPS Press, 2013, pp. 99–104, ISSN: 2308-3484, ISBN: 978-1-61208-310-0, URL: http://www.thinkmind.org/download.php?articleid=infocomp_2013_5_10_10114 [accessed: 2014-11-30].

[41] P. Leitão, U. Inden, and C.-P. Rückemann, "Parallelising Multi-agent Systems for High Performance Computing," in Proceedings of The Third International Conference on Advanced Communications and Computation (INFOCOMP 2013), November 17–22, 2013, Lisbon, Portugal. XPS Press, 2013, pp. 1–6, ISSN: 2308-3484, ISBN: 978-1-61208-310-0, URL: http://www.thinkmind.org/download.php?articleid=infocomp_2013_1_10_10055 [accessed: 2014-11-30].

[42] "GMT - Generic Mapping Tools," 2014, URL: http://imina.soest.hawaii.edu/gmt [accessed: 2014-11-30].

[43] International Commission on Stratigraphy, "International Chronostratigraphic Chart," 2014, URL: http://www.stratigraphy.org/ICSchart/ChronostratChart2013-01.pdf [accessed: 2014-11-30].

[44] C.-P. Rückemann, "Integrating Information Systems and Scientific Computing," International Journal on Advances in Systems and Measurements, vol. 5, no. 3&4, 2012, pp. 113–127, ISSN: 1942-261x, LCCN: 2008212470 (Library of Congress), URL: http://www.thinkmind.org/index.php?view=article&articleid=sysmea_v5_n34_2012_3/ [accessed: 2014-11-30].

[45] "LX SNDX, a Soundex Module Concept for Knowledge Resources," LX-Project Consortium Technical Report, 2014, URL: http://www.user.uni-hannover.de/cpr/x/rprojs/en/\lattenkreuZ{} [accessed: 2014-11-30].

[46] C.-P. Rückemann, "Computing Optimised Result Matrices for the Processing of Objects from Knowledge Resources," in Proceedings of The Fourth International Conference on Advanced Communications and Computation (INFOCOMP 2014), July 20–24, 2014, Paris, France. XPS Press, 2014, iSSN: 2308-3484.

[47] "Multilingual Universal Decimal Classification Summary," 2012, UDC Consortium, 2012, Web resource, v. 1.1. The Hague: UDC Consortium (UDCC Publication No. 088), URL: http://www.udcc.org/udcsummary/php/index.php [accessed: 2014-11-30].

[48] "Creative Commons Attribution Share Alike 3.0 license," 2012, URL: http://creativecommons.org/licenses/by-sa/3.0/ [accessed: 2014-11-30].

[49] M. My, Pompeii Reconstructed (DVD). ArcheoLibri, produced by MyMax, in: Bonaventura, Maria Antonietta Lozzi (2007): Pompeii Reconstructed, 2007.

[50] Guardasole, Vesuvio 1270 m. Guardasole SRL, Napoli, Via Argine, 313, Italia, 2013, Postcard, 40067, Description: 1944 eruptions and present day crater, Collection: LX, Provider: BGS, Entry date: 2013.

[51] M. A. L. Bonaventura, Pompeii Reconstructed. ArcheoLibri, 2007, ISBN: 978-88-95512-23-5.

[52] C. G. Newhall and S. Self, "The Volcanic Explosivity Index (VEI): An Estimate of Explosive Magnitude for Historical Volcanism," JGR, vol. 87, 1982, pp. 1231–1238.

[53] Achievements and Successful Solutions with Big Data and Computing Challenges: National and International Perspectives. International Expert Panel, July 23, 2014, The Fourth International Conference on Advanced Communications and Computation (INFOCOMP 2014), DataSys, July 20–24, 2014, Paris, France, 2014, URL: http://www.iaria.org/conferences2014/filesINFOCOMP14/INFOCOMP2014_EXPERT_PANEL.pdf.

[54] C.-P. Rückemann, "Big Data – Bigger Knowledge – Biggest Cognition: Optimising Organisation & Structure for Exa-Scale," International Expert Panel on Achievements and Successful Solutions with Big Data and Computing Challenges: National and International Perspectives, July 23, 2014, The Fourth International Conference on Advanced Communications and Computation (INFOCOMP 2014), DataSys, July 20–24, 2014, Paris, France, [Lecture], 2014, URL: http://www.iaria.org/conferences2014/filesINFOCOMP14/INFOCOMP2014_EXPERT_PANEL.pdf [accessed: 2014-11-23].

[55] "Google," 2014, URL: http://www.google.com [accessed: 2014-11-30].

Proxemic Interactions with Multi-artifact Systems

Henrik Sørensen and Jesper Kjeldskov
Research Centre for Socio-Interactive Design
Department of Computer Science
Aalborg University, Denmark
{hesor, jesper}@cs.aau.dk

Abstract— **The artifact ecologies emerging from an increasing number of interactive digital artifacts, capable of communicating with each other wirelessly, have created an interaction space where software applications are no longer limited by the physical boundaries of a single device. With the new opportunities follows an added complexity that interaction designers need to address. Previous work have shown the potential of proxemic interactions as one way of dealing with design challenges of ubicomp systems. However, the work focused on interactions involving multiple digital artifacts is limited. In this paper, we analyze two multi-artifact systems from our prior work within the domain of music consumption and identify four concepts of multi-artifact interactions: Plasticity, migration, complementarity, and multi-user. These concepts forms the basis for a discussion on the potential use of proxemic interactions in the design of multi-artifact systems.**

Keywords- artifact ecology, multi-artifact systems, proxemic interactions, music systems.

I. INTRODUCTION

The establishment of a wireless network infrastructure surrounding us introduces an easier connectivity between different digital devices. In addition, to enable data sharing and synchronization it provides great potential for interactions transcending the physical boundaries of individual devices [1]. Jung et al. [2] describe this network of devices as a personal ecology of interactive artifacts and defines it as *"a set of all physical artifacts with some level of interactivity enabled by digital technology that a person owns, has access to, and uses".* Taking advantage of the potential offered by artifact ecologies without introducing additional complexity to the user is however a challenge.

Interaction designers have become quite good at designing desktop applications and are in a post-desktop era progressively getting better at designing mobile applications as well. However, it is our belief that good interaction design for artifact ecologies consists of more than the aggregation of good designs for each individual artifact. Previous work has already moved towards an understanding of the composition [2] and dynamics [3] of the ecologies as a whole. What we find is that there is a gap between the work on understanding interactions with single artifacts and understanding personal artifact ecologies on a high abstraction level. There seems to be a challenge in understanding the interaction with multi-artifact systems that combine artifacts from personal artifact ecologies. This creates an additional layer of complexity that requires us to think of these sub-systems in a holistic way on an abstraction layer above the single artifact but below the entire artifact ecology.

The idea of proxemic interactions is to take advantage of the significance of spatial organization to the way we interact with people and digital artifacts. This has shown a great potential in helping us understand the artifact associations that constitutes multi-artifact systems and help us facilitate the interaction with them. The concept of proxemic interactions caters very well to the flexible, mobile, and wireless nature of the systems and removes some of the responsibility of handling the added complexity of multiple artifacts being in play simultaneously.

The overall goal is to move towards multi-artifact interaction designs that deliberately exploit the synergetic effects that emerge from artifact compositions and take advantage of the new opportunities this gives us without compromising user experience. The contribution of this paper is to identify concepts of multi-artifact systems that we find to be of particular significance to an artifact ecology context and explore proxemic interactions [4] as a possible framework to reveal opportunities and address design challenges for each of the identified concepts. The analysis is based around multi-artifact systems from our previous work in the music consumption domain.

Our focus lies in the interaction between humans and artifacts on a conceptual level, although it is clear that interaction designs spanning multiple artifacts is highly dependent on a comprehensive and flexible technical infrastructure for artifact discovery, connection, and communication. Therefore, we work under the assumption that this is or will be available to some extent, but acknowledge that some of the challenges are in the interaction itself.

First, we present related work on artifact ecologies, proxemic interactions, and music consumption in Human-Computer Interaction (HCI). We then clarify our understanding and delimitation of the multi-artifact system concept followed by a description of the two music systems from our prior work. Finally, we analyze the systems to identify characteristic concepts of multi-artifact systems and discuss the application of proxemic interaction before we conclude the paper with implications for future work.

II. Related Work

This section relates our work to previous research in artifact ecologies, proxemic interactions and music consumption.

A. Artifact Ecologies

In a study of the social role of products, Forlizzi [5] introduces a product ecology framework used to describe the dynamic aspects of use. The framework puts the product in the middle, meaning that each individual product has its own ecology in which components are interconnected. A product for example often has certain relations to other products that together act as a system. The components included in the framework, besides other products, are people, activities, place, and the routines and cultural context. Forlizzi's product ecology framework provides means to reason about the single product and its social impact across users.

Artifact ecologies represent a different approach of putting an ecological thinking into play in relation to the products surrounding us. Jung et al. [2], places the user in the center and define a personal ecology of interactive artifacts that a person owns, has access to, and uses. This means that an ecology is defined from the perspective of a person instead of a product/artifact. In their work, they conducted two types of exploratory studies with the common goal of understanding the relationships within artifact ecologies. Their study works under the assumption that the experience with an artifact can only be fully understood when it is considered in relation to an artifact ecology. We find the personal perspective very useful in understanding interactions that involve several artifacts. The limitation of the framework is that it does not take into account what happens when different personal ecologies intersect in multi-user interactions.

Jung et al. [2] argues that artifact ecologies are dynamically evolving. Bødker and Klokmose [3] follow up on that idea and emphasize the importance of not only understanding a current composition of artifacts in our surroundings but also how relationships among them change over time. Using Activity Theory as their theoretical framing and the Human-Artifact Model [6] as an analytical tool, they identify three states of an artifact ecology: The *unsatisfactory*, the *excited*, and the *stable* state. The artifact ecology of a person will change state over time and at some point reach the unsatisfactory state once again. Changes to the ecology can then put it into an excited state and the cycle repeats itself. One challenge they encountered in their analysis was to describe what the artifacts of artifact ecologies is. While Jung et al. [2] describes artifacts as physical objects, Bødker and Klokmose [3] found from their study that this does not always tell the whole story and that something more may be needed to systematically address the software as well.

B. Proxemic Interactions

According to Hall [7], interactions between individuals are highly influenced by interpersonal distance. There is for example a significant difference in how we interact with a person standing right in front of us compared to a person we see from across the street. A noteworthy contribution of Hall's work is the definition of discrete proxemic zones surrounding us, called the *intimate*, *personal*, *social*, and *public* zone. Each characterizes the interaction with people in our surroundings based on the immediate distance.

Vogel and Balakrishnan [8] adopts this notion in their framework for shareable interactive ambient displays and uses it to define what they refer to as *interaction phases*. This is a very direct interpretation of Hall's proxemic zones [7], which allow a large display to adapt to the user, based on distance in much the same way as people adapt to other people approaching. Greenberg et al. [4] have later expanded on the idea of proxemics as a means to describe relations in small space ubicomp environments that include multiple users, devices, and non-digital features. In their framework, they operationalize proxemics through five dimensions of proxemic interactions: *Distance*, *orientation*, *movement*, *identity*, and *location*. In addition to the theoretical framework on proxemic interactions, Marquardt et al. [9] have presented a proximity toolkit, which gives developers and interaction designers easy access to a prototype environment for proxemic interactions. The toolkit has been used in the development of prototypes such as the Proxemic Peddler [10].

Although previous work have established a deeper understanding of proxemic interactions and the potential of the framework over the last few years, Marquardt and Greenberg [11] notice that little work is applying the theory to interaction designs in ubicomp research. In their work, they elaborate on how proxemic interactions can address particular challenges of ubicomp interaction design. The six core challenges they identify in relation to proxemics are *revealing interaction possibilities*, *directing actions*, *establishing connections*, *providing feedback*, *avoiding and correcting mistakes*, and *managing privacy and security*.

Proxemic interaction shows great potential, but also comes with a risk. Because proxemic interaction relies on close tracking of people and devices, it comes with the risk of being exploited. Greenberg et al. [12] identifies so called *dark patterns* of proxemic interactions and discuss the framework from a critical perspective. A particular challenge of systems that base decisions on implicit interactions is for example to design ways for the user to opt out of the interaction. The point of context awareness in general is that the system becomes able to infer how a user wants to interact with devices based on context. The problem is when the interaction designer is not using that information in the best interest of the user but for instance in the interest of a company that wants to sell a product.

C. Music Consumption

Music has always been an important domain across disciplines due to its universal appeal. Holmquist talks about the field of ubiquitous music and how it formed through a digitization of music, portable music players and heightened bandwidth [13]. Although the article is from 2005, ubiquitous music has only become more relevant after the emergence of Internet streaming services and affordable multi-room music systems. However, Liikkanen et al. [14] point out that music consumption as a defined area in HCI research is extinct. They argue that research on music consumption through interactive devices continues but is marginal and needs a revival.

An aspect of music consumption with a particular relevance to our context is the role of music in a social setting. Leong and Wright [15] found that the increasing connectivity of technologies we use to consume music have prompted users to create their own configurations that allows them to obtain more meaningful social interactions through music. They comment on the future designs of music discovery beyond virtual social networks that utilizes the physical environment. Capital Music [16] is an example that allows co-located strangers to share music recommendations. Their study shows how music can influence social interactions in public spaces without people listening to music together. Another study explores this premise in tuna [17], which allows co-located users to "tune" in to other people's mobile music player.

O'Hara and Brown's [18] book contains a large collection of contributions to the social aspect of music consumption. Their work provides valuable insights into the sociality of music but is also a testament to how the technologies involved in music consumption has changed drastically in few years. Ongoing work is similarly exploring shared music experiences supported through technology and novel interaction designs. An example is Mo by Lenz et al. [19], which is a music player with an integrated speaker focusing on a shared music experience. Mo can be brought into a social setting and creates a connected music system by placing it next to other players.

III. Multi-artifact Systems

Before we start conceptualizing multi-artifact systems in artifact ecologies, it is important for us to clarify what we mean by the term in the first place and how we delimit it to reflect our scope. By multi-artifact systems, we refer to interactive systems, which are part of an artifact ecology, and involves more than a single physical artifact. Different terms have previously been used to describe similar concepts. Rekimoto has for instance described it as multiple-computer user interfaces with a focus on graphical user interfaces [20]. Furthermore, Terrenghi et al. [21] created a taxonomy for what they refer to as multi-person-display ecosystems, and Anzengruber et al. [22] similarly talk about display ecosystems as the platform for social feedback. As much as we appreciate the desirable features of the visual aspect, we also want to acknowledge other modalities of input and output, especially since our point of departure is in the music domain. Because we want to continue the ongoing work on artifact ecologies, it makes sense to refer to the sub-sets of artifacts as multi-artifact systems. According to the systems' view, the essential properties of an organism, or a system, is the properties of the whole that none of the parts has alone [23]. This view fits perfectly well with our intention of addressing interaction design for systems, which provides more than cross-platform interfaces.

A. Delimitation

In our definition and delimitation of multi-artifact systems, we acknowledge Bødker and Klokmose's [3] comment on the artifact term encompassing more than the physical interactive artifact. Our interest lies in the interaction designs, which transcends the boundaries of a physical artifact, thus we use multi-artifact systems as a term to describe sub-systems of artifact ecologies consisting of a specific composition of hardware *and* software artifacts used throughout a particular activity. This could technically involve the interaction with a desktop-PC communicating with a web server through a browser, but our work specifically aims at systems where either the user provides direct input to the artifacts or the artifacts provide direct output to the user. The server part of the example fulfils neither role. Video conferencing is another example that involves several artifacts, but traditionally only one from each user's personal ecology. It is thus not a multi-artifact system either. A system that merges persons' smartphones into a common interface is an example of a multi-artifact system from our perspective, as it would become a multi-artifact system in each user's ecology. The example shows the inclusion of systems that exist in the intersection between personal artifact ecologies, where multiple persons interact with some or all of the same artifacts.

B. Time and Space

Although the browser and video conferencing examples provide some limitation to our scope it should not be interpreted as if the artifacts in the multi-artifact systems are required to be co-located or that the interaction with each artifact has to happen simultaneously. We still consider systems that distribute interaction across time and space as long as the interaction is part of the same activity from the personal point-of-view. The important point is that the system provides more than a cross-platform interface. An example is the Google Chrome browser. Having a version for Windows, Android and iOS makes it a collection of single-artifact interactions, but when it starts remembering tabs, bookmarks, search preferences, etc. across artifacts it becomes interaction with a multi-artifact system.

The following sections provide descriptions of the two multi-artifact music systems from prior work, on which we base our conceptualization.

IV. AIRPLAYER

AirPlayer is a multi-room music system that adapts to the location of the user with the purpose of creating an implicit control of the music. It consists of a .NET C# server application and an Android client that builds on top of Apple's AirPlay protocol stack, hence the name, making it capable of streaming music from a central digital music library to speakers placed in different locations around the home. Each speaker connects wirelessly to a central music player application through an Airport Express that doubles as a Wi-Fi access point. The use of a Wi-Fi network additionally makes it possible for the user to control music independently in specific locations from a smartphone application. AirPlayer handles separate locations through the notion of *zones*. A zone is per default a representation of the room in which a particular Airport Express is placed. However, the user can combine zones to play and control music in several locations simultaneously. The zone name is visible in the bottom of the main screen (see Figure 1) and by sliding horizontally, the user can manually cycle through the different zones to see the current song playing change the volume etc.

Similar features are already present in Apple's existing product family, through iTunes, as well as in other multi-room music systems. The significance of AirPlayer is its addition of proxemic interaction features that allow the system to adapt to spatial relations between the user and particular speakers placed in different rooms. The proxemic interaction manifests itself in AirPlayer as two features called *location* and *movement*, which the user enables through the smartphone application. A simple implementation of an indoor positioning system provides the necessary logic to estimate user location and distance to individual speakers. The smartphone application continuously measures Received Signal Strength Indicator (RSSI) values from the Airport Express Wi-Fi access points and uses them to determine in which zone the user is located. Chen and Kobayashi [24] argues that indoor location is feasible through radio signal based indoor location, given an implementation of a sound method for signal propagation. Although the proxemic sensing in this prototype is not based on a sophisticated algorithm for signal propagation, in practice it performs to a degree that enables the user to experience proxemic interaction.

A. Location

When the location feature is enabled, the smartphone application continuously adapts to represent the music currently playing in the zone where the user is located. As illustrated in Figure 1, this means that the user interface presents information about the song playing and ensures that the user automatically controls the music in this particular location. The change happens in a seamless and subtle way, when the user changes location, without the need for explicit user interaction. Whenever the system detects a change in location, it simply adapts the smartphone application to represent the current zone. From the user point-of-view the interaction is similar to having a universal remote control that can be used to control independent music players in each room.

B. Movement

When the movement feature is enabled, music follows the user around the home as illustrated in Figure 2. By tracking the smartphone, the system is able to anticipate which zone the user is entering, continue the music in the new zone, and stop the music in the old one. What actually happens is that AirPlayer streams the music to all zones simultaneously and adjusts the volume in accordance to the movement of the user. The movement and location feature can be enabled independently but are not strictly independent of each other. When the movement feature is enabled, so is the location feature as the same music is always playing where the user is located. The location feature enables a state where the smartphone user interface adapts to the location of the user and the music content stays. Inversely, the movement feature enables a state where the user interface stays the same and the music content adapts to the location of the user.

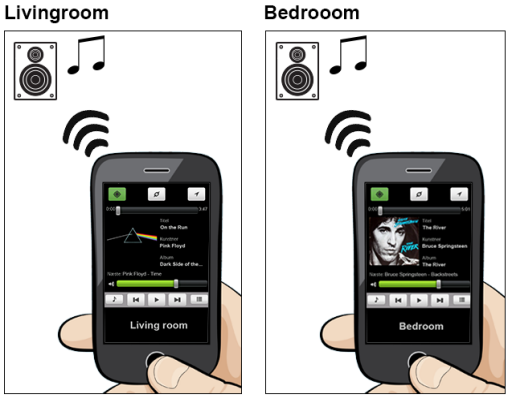


Figure 1. The location feature adapts the user interface and control to the location of the user.

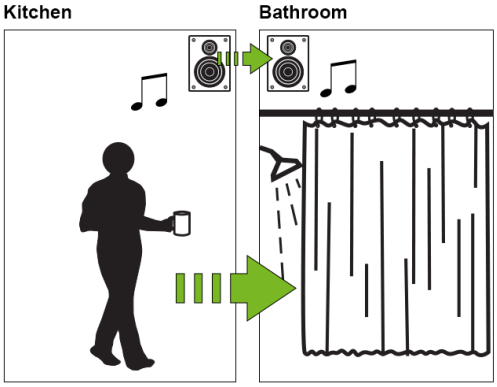


Figure 2. The movement feature makes music follow the user across locations.

V. MEET

The second system, called Music Experienced Everywhere Together (MEET), is a multi-user, multi-artifact music system that addresses the problem domain of playing music in a social setting where there is no DJ or other dedicated control of the music. The concept of MEET is to allow several co-located users to share their music to a music player at social events, thereby creating a common music library. What is being played from the library is then controlled in a collaborative manner where anyone can nominate and vote for songs using their own smartphones or a dedicated tablet. To nominate a song simply means that the system puts a song up for display as a possible song to be played next. When it will be played, or whether it will be played at all, is up to the crowd and how participants choose to place their votes. The smartphone and tablet application is implemented in Android and the library and music player is implemented in Java/JavaFX. The Real-time Transport Protocol (RTP) and Real Time Streaming Protocol (RTSP) is used for streaming between library and music player.

The intention of a system like this, compared to for example a traditional jukebox, is to make it a social experience that tries to be fair to the users and that allows people to engage in the music control in a different way. The advantages of the approach are:

- No one can interrupt what is currently being played in order to put on another song.
- There is no playlist queue, but rather a system that dynamically changes to reflect what people wants to hear in the moment.
- The music in the library is not a large generic collection but a personalized collection of people's own music.

A quality of music is its ability to be experienced as a background activity. Consequently, it is important for MEET to allow participation on different levels and not become the event itself. To accommodate this requirement, MEET has a built in feature that automatically nominates songs from the music library if there are less than three current nominations.

A. Smartphone Application

The smartphone application is the primary input artifact for the music player. Besides being the interface to share music to the music player, it features a nominate functionality, where users can browse the collection of music shared by users and nominate songs they would like to hear. Furthermore, the interface presents the list of nominations, giving the option to give a positive or negative vote for nominations. Each vote will simply add or subtract one point from the total score. An important quality is to utilize the users' own smartphones, thereby making it a personal artifact representing the specific owner's choices.

B. Tablet Application

The tablet application is a simplified version of the smartphone application that only works for nominating and voting. It primarily serves as a public input artifact used by people without a compatible smartphone and secondly as a physical interaction point for the music system in general. Because the tablet is an artifact shared among several users, the vote feature is modified to allow an infinite number of votes for a single nomination and instead introduce a 10-second countdown after a vote, where the application locks itself. The lock mechanism is added in an attempt to prevent a person from exploiting the tablet by voting repeatedly for the same song.

C. Situated Display

The situated display shows the primary visual output of the music player to the users. The interface is suitable for a large flat screen TV or projector and should be placed with visibility in mind as it represents the current state of the music player to the users. An album cover represents a nomination on the situated display (see Figure 3). Size of nominations indicate score, meaning that the largest are more likely to play next. This score does not map to the smartphone application, thus the situated display is the only place where the status is visible. Figure 3 shows the voting interface of the different artifacts.The music system is running in one place and distributes interaction to other artifacts. Specific artifacts consist of a device with a part of the distributed interface each with their own output screen and each serving a specific purpose.

Figure 3. The different artifacts of MEET and their respective GUIs for the voting functionality.

VI. Concepts of Multi-Artifact Systems

In this section, we use the two presented systems to identify concepts that we find meaningful in the context of multi-artifact systems. The concepts are not novel in themselves, but the contribution lies in the use of them as tools to describe interaction across artifacts, which can inform a focused and structured effort in the design of proxemic interactions.

A. Plasticity

The term plasticity is inspired by neuroscience and the way our brain is able to change as a reaction to external influences such as changes in the environment. The term has been adapted to HCI to describe a similar behavior for non-static user interfaces. Balme et al. [25] define plasticity applied to HCI as *"…the capacity of an interactive system to adapt to changes of the interactive space while preserving usability"*. Changes of the interactive space can both be in terms of the physical environment, the resources available or virtual changes. Plasticity is meaningful for different types of artifacts. A smartphone application can for instance adapt to the location of the user (Figure 4), or a public display can adapt to the time of day or number of people in front of it.

In AirPlayer, the location feature enables the smartphone application to adapt to the location of the user, providing information about the music currently playing, as well as control of the music in this particular location. In AirPlayer, it is the spatial relations between the smartphone and speakers placed around the home that determines what is presented to the user, which is why we argue that plasticity also has its place as a concept of multi-artifact systems.

MEET does not have any plasticity integrated in the interaction design. Each artifact has a form that plays a specific role in the system. An idea of introducing it into the smartphone application could be to provide more feedback on the status nominations, if the user is not able to see the situated display.

Another interesting challenge of artifact ecologies is the increase in general-purpose artifacts capable of executing different sort of applications. Our phone is no longer just for making phone calls, our TV is no longer just for watching TV, and the newest addition to our ecologies is smart watches that does much more than showing the time. As our collection of general-purpose artifacts expand, arguably so does the number of multi-artifact systems and thereby the complexity of them. In AirPlayer, the smartphone application adapts to contextual information within the user's current activity. Artifacts able to adapt to fit a certain activity and composition of artifacts could be an interesting utilization of plasticity.

B. Migration

Migration refers to the capability of moving the interaction from one device to another while preserving the state (see Figure 5). Berti et al. describes migratory user interfaces as *"…interfaces that can transfer among different devices, and thus allow the users to continue their tasks…"* [26]. The essential concept here is the continuity in the interaction and it is where it differs from cross-platform applications, which merely presents the user to an alternative version of the same application on different physical artifacts. In the taxonomy of migratory interfaces Berti et al. [26] distinguish between different degrees of migration: Total migration, is where the entire interface migrates from one artifact to another. In partial migration, only a part migrates to the target artifact. Distributing migration is where the interface migrates to multiple target artifacts. Finally, aggregating migration, is where the interface migrates from multiple artifacts into one.

The movement feature of AirPlayer makes music follow the user around by moving music output from one artifact to another. The interesting thing about the movement feature of AirPlayer is not that it plays the same music from a central source. It is the ability to do so continuously across locations as the user moves around. In the AirPlayer example, it is the content (music) migrating between exactly two artifacts. The way it works in AirPlayer is an example of interface migration not necessarily being a matter of transferring an application state.

Migration and plasticity are somehow related concepts that both encourage more flexible and adaptable relations in artifact ecologies. There is no implementation of interface migration in MEET but is in a similar way as plasticity a concept that could be integrated. A possible use is to transfer the state of the situated display to a view on the smartphone application as soon as the situated display is no longer visible to the user.

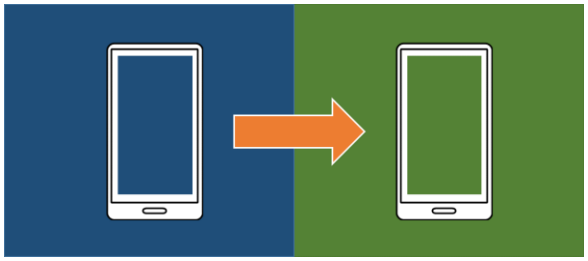


Figure 4. Plasticity allows user interactive systems to adapt to the interaction space.

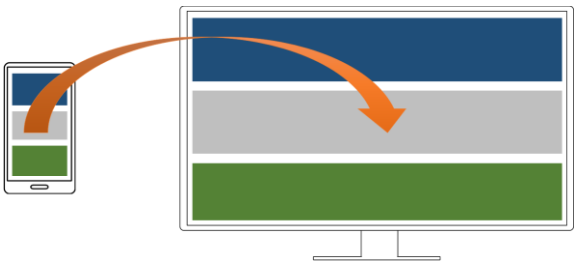


Figure 5. Migration allows interaction to move between devices.

C. Complementarity

Complementarity, as illustrated in Figure 6, is the concept of distributing a user interface across artifacts allowing simultaneously use in a collaborative fashion. A typical example is a remote control where the user input is clearly separated from the output, and one artifact is controlling another. Another type of complementary interactions have started to emerge in the form of so-called *companion apps* or *second-screen apps*, which is a mobile application that complements the interaction of another artifact. An example is an application that shows information for a TV show.

In MEET, interaction is distributed across different artifacts. The different artifacts can be described as being *complementary* to each other, as each of them provides features that improve the overall system. The music player is useless if no one has connected a smartphone, shared some music and nominated at least one song. The smartphone application similarly does nothing on its own. Distributing functionality is of course a conscious design choice that is not strictly necessary to play music at a party. However, the distribution takes advantage of available interaction resources to create a different kind of experience. What field studies of MEET have shown is also that such systems can provide an opportunity for a different social interaction and utilization of the environment, than a traditional music system. Unfortunately, the benefits come with the cost of an additional level of complexity, both technically and in the interaction design.

The complementarity between the smartphone/tablet and situated display in MEET is similar to the notion of *coupled displays* [21] where lessons can be learned from previous work. In addition, it is important to consider other modalities of input and output of multi-artifact systems than the visual, as artifacts may be able to utilize these to complement each other in different ways.

AirPlayer similarly has an element of complementarity in its interaction design although more subtle than in MEET. The smartphone application provides the input and output to a music system distributed throughout the home that provides the music output. Although the smartphone application is able to control various music outputs independently, the complementarity in AirPlayer is basically a remote control metaphor. In a way this is also the case in MEET although both examples illustrate that complementary artifacts can be more powerful than a direct mapping of a traditional remote control.

It is reasonable to talk about dependency of the relationships between complementary artifacts. In MEET there is a very strong dependency between the smartphone application, the music player, and the situated display as none of them can work independent of the other. An exception is the tablet, which can be removed without losing crucial functionality but does nothing on its own. In AirPlayer, there is similarly a strong dependency between artifacts as no control of the music is implemented outside the smartphone application. The point is that it can be useful to consider the dependencies of complementary artifacts. Not only in the scope of the multi-artifact system but also in relation to the artifact ecologies involved. In AirPlayer all the artifacts belongs to the ecology of a single person as only one smartphone application is allowed at any time. MEET on the other hand is by design dependent on artifacts from several personal artifact ecologies.

D. Multi-user

Multi-user interaction is quite self-explanatory and is simply the concept of interactions that involve more than one user. However, it is worth making the distinction between two cases. One is where multiple users interact with a system simultaneously (Figure 7). The other is where multiple users interact with a system one at a time.

The multi-user concept is different from the others, as it addresses the users instead of the artifacts. Whether a system is designed for single or multi-user interaction is not surprisingly an important factor. What it means to include multiple users in terms of artifact ecologies is that the multi-artifact system spans more than one personal artifact ecology and that all involved users' ecologies intersect.

MEET is for instance designed specifically for a social context with several simultaneous users. Each user's smartphone is a part of their individual artifact ecology and can serve various purposes in different contexts. When they arrive and connect their smartphone to the player the situated display and music player becomes part of each user's artifact ecology as well. Even though smartphone at this point is part of the same multi-artifact system, they are not part of any other user's artifact ecology.

Figure 6. Complementarity allows interactions to be distributed across devices.

Figure 7. Multi-user interactions are either interactions involving more than one user simultaneously or one at a time.

The new possibilities for designing multi-user interactions is one strength of multi-artifact systems. MEET for example, has no inherent upper limit on the number of simultaneous users by design. The possibilities do however come with a price. Just as multi-artifact systems adds an extra layer of complexity to single-artifact interaction, so does multi-user interaction. It is interesting to see how some multi-artifact systems are inherently designed for a single user but where it is trivial to support more simultaneous users. AirPlayer, on the other hand, is a case where it easily gets complicated if it needs to support more user even though it would make sense in an everyday situation. MEET is specifically designed to support simultaneous users and would simply be a different system if it were to support a single-user mode.

E. Comparing AirPlayer and MEET

We have analyzed the two multi-artifact music systems, MEET and AirPlayer and have identified four concepts of multi-artifact interactions: *Plasticity*, *migration*, *complementarity*, and *multi-user*. In AirPlayer we identified the concepts of plasticity, migration, and complementarity and in MEET we identified complementarity and multi-user (Figure 8). Complementarity was the only overlapping concept and served a similar purpose in both systems, namely to distribute part of the user interface onto a smartphone. A difference is that in AirPlayer there was no other visual interface besides the smartphone application. The multi-user concept differs from the others, as it does not refer to the artifacts. It is therefore interesting to see how important it is to the way a multi-artifact system is designed.

We do want to stress that the concepts are not individual solutions to multi-artifact interaction design. There lies great opportunity in combining the concepts as was also evident in our analysis. Plasticity, migration and complementarity in AirPlayer serves a particular purposes for a part of the system and a strength of the combination can be seen in the movement feature. Having the music follow you around could be achieved by simply playing it from the smartphone itself, but the music system installed in the home is of a much higher quality and through migration, music can still follow the user around. The convenience of controlling music on the smartphone, offered through the complementary interface, is on the other hand preferable. Partial, distributing, and aggregating migration can be used to switch between complementary artifact compositions.

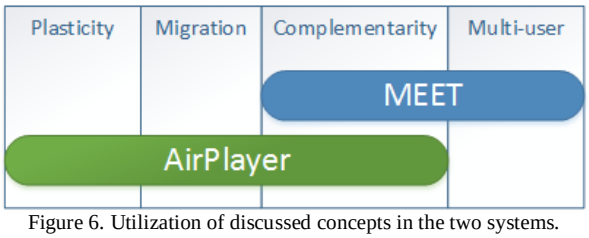


Figure 6. Utilization of discussed concepts in the two systems.

VII. Proxemic Interactions

In this section, we discuss the use of proxemic interactions as a possible interaction framework for multi-artifact systems in artifact ecologies. We specifically revisit the four identified concepts of multi-artifact systems described earlier, and discuss possible opportunities and challenges in the application of proxemic interactions to realize them. The discussion is based on our experiences from the studies of the music systems as well as insights from previous work on proxemic interactions.

A. Proxemics and Plasticity

Plasticity represents the very basic concept proxemic interactions was defined for: Adapting the user interface of interactive systems to better accommodate the spatial organization of people, digital artifacts and non-digital objects. This is what Vogel and Balakrishnan [8] demonstrates in their work on a public display that adapts to the distance of a user in a seamless way through four interaction phases very analogous to Hall's [7] proxemic zones. Similar work on proxemic interactions focuses on the distance between a user and a large display, and there is a great potential in the use of adapting user interfaces of artifacts based on the proxemic relations to nearby users.

In AirPlayer, we saw how plasticity was used to adapt the content of the smartphone interface according to the *location* of the user. A noteworthy detail here is that it is the content that changes and not the state of the user interface. Another aspect of plasticity is to allow the user interface to adapt to accommodate the surroundings. In MEET, it could for example be interesting to let the interface of the situated display adapt to the number of users in front of it. This could be used to either improve the experience of current users or help attract more.

A general challenge with proxemic interactions as a way of automating plastic user interfaces is the dilemma of how much the user needs to understand the decisions made by the system. A smartphone typically uses a proximity sensor to disable the touchscreen when a call is picked up and the phone is being held in a position close to the user's ear. Not everyone knows this happens and as long as it prevents accidentally pressing unwanted buttons, the feature serves its purpose. In the design of proxemic interactions that adapt the user interface, it is however important to take into consideration how much the user is kept in the dark.

B. Proxemics and Migration

The idea of interface migration is very relevant in an artifact ecology context as we already own and interact with several artifacts capable of performing the same tasks. Without some sort of preserved state across artifacts, we end up restarting interactions every time we switch between them. Which artifact is appropriate for the task in a given point of time depends on various factors and presents a challenge that does not seem to be completely solvable by proxemic interactions alone.

In AirPlayer, the music migrates from artifact to artifact, depending on the movement of the user, i.e., the content. This example already shows how it can make sense to base the migration on proxemic relations. The implementation of proxemic interactions in AirPlayer is rather coarse-grained and only works on a room level. The *location* dimension in proxemic interaction theory differs from *distance* in that other features of the location can be significant. In the AirPlayer example, the type of room could for example be meaningful to the decision about migrating and if the user left the house, it could make sense to perform a total migration to the smartphone.

More generally, proxemic relations seems like a natural approach to interface migration. Many of the challenges identified by Marquardt and Greenberg [11] apply to interface migration such as revealing migration targets, directing actions and establishing connections. A general challenge that however also would apply to migration is how to opt out or how to avoid automatically opting in. It is easy to assume that the user would always want to migrate the current task to the nearest and/or best artifact. However, there could be situations where this is not the case. It would for instance not be appropriate to migrate mobile internet browsing to every public display a user passes by even though it provides a larger screen.

C. Proxemics and Complementarity

Unlike plasticity and migration, complementarity as a concept does not infer any ability to adapt or change the interface of an interactive system. It rather describes how artifacts can complement each other to allow for an augmented interaction experience. The concept is still relevant to discuss in relation to proxemic interactions as the procedure of connecting artifacts and making meaning of current associations is not a trivial task.

Results from the field study of the complementary interface of MEET shows the importance of the spatial organization of users and artifacts in the physical environment. In cases where the user had great visibility of the situated display, the most important property was the coordination of visual feedback between the situated display and the smartphone application. However, when either the users were at a distance or otherwise unable to see the display clearly, they would be highly dependent on the limited feedback given from the smartphone application. It can be argued that a redesign of the interface or added features would solve this problem, but an important aspect of the smartphone application is also the simplicity as the users were engaged in a social activity as well.

In an artifact ecology context, there generally seems to be an unlocked potential in utilizing proxemic interactions to combine plasticity with complementarity. Mobile artifacts serve multiple purposes that often overlap. Configuring the roles of artifacts in our immediate surroundings is currently up to the user and as the number of artifacts grow, it becomes a difficult task to get the best out of the artifacts in a given situation. Here we see a potential for proxemic interactions to adapt the interface of the individual devices to complement other artifacts in its proximity. The limitation of proxemic interactions in relation to plasticity is that spatial relations do not uniquely characterize an activity. The couch in front of the TV can be the place where a user watches movies while using a smartphone as a remote control, but it might as well be where he takes a nap.

D. Proxemics and Multi-user Interactions

Supporting multiple co-located users in multi-artifact systems is far from a trivial challenge. The *identity* dimension of the proxemic interaction framework do acknowledge the importance of distinguishing between users. This is similar to how we might feel more comfortable having a conversation very close to our spouse than to a stranger, and in the interaction design of multi-artifact systems this makes sense as well. Designing proxemic interactions based on the identity of multiple users is very useful and can help in managing privacy and security through proximity-dependent authentication [11]. A laptop should, e.g., react differently if it is aware that the owner is sitting in front of it with a smartphone than if it is an unauthorized user. However, there are other underlying challenges of proxemic interactions in multi-user scenarios.

Commercial systems heavily rely on a model similar to the artifact ecology with a single user in the center. Everything is built around user profiles, which inherently are meant for one user at a time. The problem is that it is not always obvious what it means to support multiple simultaneous users. The idea of the movement feature in AirPlayer, where music follows you around is an example that makes perfectly good sense for one person. It is however difficult to design appropriate behavior if more people want to use the feature simultaneously. What happens if two persons, with different music following them, enter the same room? Rules could be defined to cope with this specific problem, but what could be more interesting is to explore generic approaches. As it may seem trivial to take the number of intended users into account for a particular context, we find that existing solutions shows it is an important area to do more work to understand the multi-user dynamics of artifact ecologies.

VIII. Conclusion and Future Work

The work in understanding artifact ecologies becomes important as the number of relationships among artifacts increase in complexity. What we have done is to start an articulation of the sub-systems of artifact ecologies on a level in between the interaction with single artifacts and the understanding of the ecologies in their entirety. The four identified concepts of multi-artifact systems, i.e., *plasticity, migration, complementarity, and multi-user* can help obtain a more fine-grained understanding of artifact ecologies, which informs a discussion of the concepts in relation to proxemic interactions.

The discussion has revealed specific pointers to proxemic interaction's potential for the design of multi-artifact systems and identified limitations of spatial relations as context. As the identified concepts are deduced from the interaction design of two multi-artifact systems, we make no claim of completeness. A next step would therefore be to get a broader understanding of interactions with multiple artifacts on a conceptual level with the goal of creating design guidelines for proxemic interactions in multi-artifact systems that do not only work well in isolation, but fits into an artifact ecology.

REFERENCES

[1] H. Sørensen and J. Kjeldskov, "Concepts of multi-artifact systems in artifact ecologies," Proc. International Conference on Advances in Computer-Human Interactions (ACHI 2014), IARIA, 2014, pp. 141-146.

[2] H. Jung, E. Stolterman, W. Ryan, T. Thompson, and M. Siegel, "Toward a framework for ecologies of artifacts: how are digital artifacts interconnected within a personal life?," Proc. Nordic Conference on Human-Computer Interaction: Building Bridges (NordiCHI '08), ACM Press, 2008, pp. 201-210, doi: 10.1145/1463160.1463182.

[3] S. Bødker and C. Klokmose, "Dynamics in artifact ecologies," Proc. Nordic Conference on Human-Computer Interaction: Making Sense Through Design (NordiCHI '12), ACM Press, 2012, pp. 448-457, doi:10.1145/2399016.2399085.

[4] S. Greenberg, N. Marquardt, T. Ballendat, R. Diaz-Marino, and M. Wang, "Proxemic interactions: the new ubiComp?," interactions, vol. 18, no. 1, Jan. 2011, pp. 42-50, doi: 10.1145/1897239.1897250.

[5] J. Forlizzi, "How robotic products become social products: an ethnographic study of cleaning in the home," Proc. ACM/IEEE International Conference on Human-robot Interaction (HRI '07), ACM Press, 2007, pp. 129-136, doi:10.1145/1228716.1228734.

[6] S. Bødker and C. Klokmose, "The human-artifact model: an activity theoretical approach to artifact ecologies," Human-Computer Interaction, vol. 26, no. 4, 2011, pp. 315-371, doi: 10.1080/07370024.2011.626709.

[7] E.T. Hall, "The Hidden Dimension," Doubleday, 1966.

[8] D. Vogel and R. Balakrishnan, "Interactive public ambient displays: transitioning from implicit to explicit, public to personal, interaction with multiple users," Proc. ACM Symposium on User Interface Software and Technology (UIST '04), ACM Press, 2004, pp. 137-146, doi: 10.1145/1029632.1029656.

[9] N. Marquardt, R. Diaz-Marino, S. Boring, and S. Greenberg, "The proximity toolkit: prototyping proxemic interactions in ubiquitous computing ecologies," Proc. ACM Symposium on User Interface Software and Technology (UIST '11), ACM Press, 2011, pp. 315-326, doi:10.1145/2047196.2047238.

[10] M. Wang, S. Boring, and S. Greenberg, "Proxemic peddler: a public advertising display that captures and preserves the attention of passerby," Proc. International Symposium on Pervasive Displays (PerDis '12), ACM Press, 2012, Article 3, 6 pages, doi:10.1145/2307798.2307801.

[11] N. Marquardt and S. Greenberg, "Informing the design of proxemic interactions," IEEE Pervasive Computing, vol. 11, no. 2, Apr. 2012, pp. 14-23, doi: 10.1109/MPRV.2012.15.

[12] S. Greenberg, S. Boring, J. Vermeulen, and J. Dostal, "Dark patterns in proxemic interactions: a critical perspective," Proc. Conference on Designing Interactive Systems (DIS '14), ACM Press, 2014, pp. 523-532, doi: 10.1145/2598510.2598541.

[13] L.E. Holmquist, "Ubiquitous music," Interactions – Ambient intelligence: exploring our living environment, vol. 12, no. 4, July + August 2005, pp. 71-ff, doi: 10.1145/1070960.1071002.

[14] L. Liikkanen, C. Amos, S.J. Cunningham, J.S. Downie, and D. McDonald, "Music interaction research in HCI: let's get the band back together," CHI '12 Extended Abstracts on Human Factors in Computing Systems (CHI EA '12), ACM Press, 2012, pp. 1119-1122, doi:10.1145/2212776.2212401.

[15] T.W. Leong and P.C. Wright, "Revisiting social practices surrounding music," Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI '13), ACM Press, 2013, pp. 951-960, doi:10.1145/2470654.2466122.

[16] J. Seeburger, M. Foth, and D. Tjondronegoro, "The sound of music: sharing song selections between collocated strangers in public urban places," Proc. International Conference on Mobile and Ubiquitous Multimedia (MUM '12), ACM Press, 2012, Article 34, 10 pages, doi:10.1145/2406367.2406409.

[17] A. Bassoli, J. Moore, and S. Agamanolis, "tunA: socialising music sharing on the move, (book chapter)" in K. O'Hara and B. Brown (eds.), Consuming Music Together: Social and Collaborative Aspects of Music Consumption Technologies, Springer, 2006.

[18] K. O'Hara and B. Brown, "Consuming Music Together: Social and Collaborative Aspects of Music Consumption Technologies", Springer, 2006.

[19] E. Lenz, S. Diefenbach, M. Hassenzahl, and S. Lienhard, "Mo. shared music, shared moment," Proc. Nordic Conference on Human-Computer Interaction: Making Sense Through Design (NordiCHI '12), ACM Press, 2012, pp. 736-741, doi:10.1145/2399016.2399129.

[20] J. Rekimoto, "Multiple-computer user interfaces: "beyond the desktop" direct manipulation environments," Proc. CHI '00 Extended Abstracts on Human Factors in Computing Systems (CHI EA '00), ACM Press, 2000, pp. 6-7, doi: 10.1145/633292.633297.

[21] L. Terrenghi, A. Quigley, and A. Dix, "A taxonomy for and analysis of multi-person-display ecosystems," Personal and Ubiquitous Computing, vol. 13, no. 8, November 2009, pp. 583-598, doi: 10.1007/s00779-009-0244-5.

[22] B. Anzengruber, G. Castelli, A. Rosi, A. Ferscha, and F. Zambonelli, "Social feedback in display ecosystems," IEEE International Conference on Systems, Man, and Cybernetics (SMC '13), IEEE, 2013, pp. 2893-2898, doi: 10.1109/SMC.2013.493.

[23] F. Capra, "The Web of Life," Anchor Books, 1996.

[24] C. Yongguang, and H. Kobayashi, "Signal strength based indoor geolocation," IEEE International Conference on Communications (ICC '02), IEEE, 2002, vol. 1., pp. 436-439, doi: 10.11.1109/ICC.2002.996891.

[25] L. Balme, A. Demeure, N. Barralon, J. Coutaz, and G. Calvary, "CAMELEON-RT: a Software architecture reference model for distributed, migratable, and plastic user interfaces," Proc. Second European Symp. Ambient Intelligence (EUSAI 2004), Springer, 2004, pp. 291-302, doi:10.1007/978-3-540-30473-9_28.

[26] S. Berti, F. Paternò, and C. Santoro, "A taxonomy for migratory user interfaces," Proc. 12th Int. Workshop on Interactive Systems. Design, Specification, and Verification (DSVIS 2005), Springer, 2005, pp. 149-160, doi: 10.1007/11752707_13.

Games as Actors

Interaction, Play, Design, and Actor Network Theory

Jari Due Jessen
Center for Playware
Technical University of Denmark
2800 Kgs. Lyngby, Denmark
jdje@elektro.dtu.dk

Carsten Jessen
Centre for Teaching Development and Digital Media
Aarhus University
2400 Copenhagen, Denmark
cj@dpu.dk

***Abstract*—When interacting with computer games, users are forced to follow the rules of the game in return for the excitement, joy, fun, or other pursued experiences. In this paper, we investigate how games achieve these experiences in the perspective of Actor Network Theory (ANT). Based on a qualitative data from a study of board games, computer games, and exergames, we conclude that games are actors that produce experiences by exercising power over the user's abilities, for example their cognitive functions. Games are designed to take advantage of the characteristics of the human players, and by doing so they create in humans what in modern play theory is known as a "state of play".**

Keywords: computer games; board games; Actor Network Theory; interaction; game research; game design; play theory

I. INTRODUCTION

Using computer software usually implies that the user is the active part who controls the interaction by input and direct manipulation [1] [2]. Interaction with computer games is a different experience because the user acts in a game world where the contents of the game has extensive influence on the gamer's behavior. Game figures and other game items are not just passive objects that can be manipulated by the gamer. For a game to come live, gamers have to follow rules and act as the game requires. Playing a computer game like *Counter Strike* [3] or *World of Warcraft* [4] is not just a question of manipulating an avatar. The game is forcing the gamer to react to events in the game by acting in a certain way if he wants to survive and prosper in the game, i.e., the gamer is placed in a role he has to fulfill. In other words: games do something to and with people who play them and, in a certain way, games are just like actors who have an agency. What this agency consists of and how it is engineered is of interest to designers.

In this article, we will demonstrate how games can be seen as actors and as organizers of actors and actions on the basis of Actor Network Theory (abbreviated to "ANT") [5]. ANT is well suited for the analysis of interaction with games by users since ANT offers an approach to agency that does not assign power only to human actors but allows the possibility for objects and rules to be examined as actors. Also, ANT opens the door to viewing design as a social enterprise. As Yaneva stresses: "…design has a social goal and mobilizes social means to achieve it" [6].

ANT has received some attention in game studies during the last decade. Several scholars have studied games on the basis of ANT [7], focusing primarily on the interchange between humans and technology [8] or on the development of social networks in online games [10]. We take a different approach and show how the ANT perspective can explain which forces are at work when games are actually played. Our focus is thus on defining the immediate effects of using games. Our approach is in line with Seth Giddings [9], who have analysed games from the perspective of ANT. Giddings stresses that "the analysis of video games […] demands the description of a special category of nonhumans, software entities ([…] agents] that act more or less autonomously or effect emergent behaviour" [9].

The article is the result of a research project where we studied gamers of different ages playing computer games, board games, and digital play equipment. Contrary to Giddings and other scholars studying computer games, our point of departure was the theory that computer games and other games based on digital technology are games before they are anything else [10]. They are not first and foremost technology. Therefore, the study is focused on studying games as a genre rather than just digital games, and our main example here is a board game.

In the next section, we will introduce ANT focusing mainly on the concept of "translation" which is employed as our main analytical foundation. After this, the paper will present the research methodology applied for collecting data. In the following sections, the selected case of game playing will be presented followed by a presentation and a discussion of the results of our investigation. In this section, we will also draw on modern play theory to further explain how and why games function and also why computer games belong to the general genre of games. We conclude this article with reflections on how our viewpoint influences design.

II. ACTOR NETWORK THEORY

ANT was first developed by science and technology study scholars Michael Callon and Bruno Latour [11] as a new approach to social theory. ANT is of interest to any analysis of technology which goes beyond the assumption that technology is a mere instrument that we, as humans, utilize. ANT claims that any element of the material and social world (nature, technology, and social rules) can be an

actor in the same way humans are. Agency is never only human or social but always a combination of human, social, and technology elements [12]-[14].

ANT is not a theory in the usual sense of the word according to Latour himself, since ANT does not explain "why" a network takes a certain form or "how" this happens [5]. ANT is more a method of how to explore and describe relations in a pragmatic manner, a "how-tobook" as Latour defines it [5], and, as such, it offers a way to describe ties and forces within a network.

The main idea of ANT is that actions always take place in interaction between actors in networks when actors influence each other and struggle for power. We usually see social interaction between humans this way, however, ANT differs from traditional social theory by stating that the actors are not only humans but can be other elements as well.

A. *The traffic example*

ANT can be difficult to grasp and even counter-intuitive [12] because it reverses our common understanding of actors and agency, e.g., when it cuts across the subject-object division underlying our thinking about the world we live in. In an attempt to clarify ANT, Hanseth and Monteiro [15] use traffic as an example to explain the implications of seeing something in the perspective of ANT. We find their example very useful in obtaining a better understanding of ANT and, hopefully, what we later have to say about what games do. The following is a short presentation of their attempt and afterwards we will use it to explain the process of translation: When you are driving in your car from one place to another, you are acting, however, your acts are heavily influenced by technology, the material world (maneuvering abilities of the car, layout of roads, traffic signs, traffic regulation, etc.), and the immaterial (traffic rules, traffic culture, etc.) and habits (your own behaviour as a driver) [15].

According to ANT, these factors (including you) all function as actors and should be understood as forces of agency in a linked network. If you want to play the game, human and non-human, technical and non-technical elements are part of the network, and none of the elements are per definition granted special power over the others [12], [15].

Expanding the thoughts of Hanseth and Monteiro, we can add that, in the traffic example, you want to move from place to place, but you are dependent upon technology and forced to act in accordance with both social rules and physical conditions. Even though you are the driver, you will clearly feel the forces of other actors when acting out the driving. For instance, the road forces you to follow a certain route, the traffic light forces you to stop and start. One can say that in order to reach your goal safely and quickly, you have to "give in" to the network and in a way "hand over" your acting power and control over the car, so that the vehicle will move in accordance with the demands of traffic network. You have to "delegate" [12] power to the traffic network, and, in return, you will reach your goal as fast and safely as possible. Of course, you are not handing over the control of yourself to the network. To delegate is more to act as prescribed by other actors. According to ANT, this is what happens in an actor-network relation.

B. *Translation*

The way delegation is done is through the process of *translation*. This process requires the actors in a network to accept roles, a worldview, rules of acting, a path to follow etc. Michel Callon [16] describes the process of translation as a process of "persuading" with four distinct phases, he calls "moments": problematization, interessement, enrolment, and mobilization. These moments are inter-related overlapping steps that describe how stable actor-networks come to be established [17]. We will introduce them briefly in the following, and later use them in our game analysis.

The first moment, problematization, is where some of the actors in the network in question bring forth a definition of the problem and present a viable solution to it for the other actors. This is also the process during which the actors' roles are defined (both human and non-human actors). To use the , example above, this is where the car and the traffic network are presented as a solution to the transport problem.

As part of the problematization process, a so-called obligatory passage point (OPP) is defined, i.e., a practicable solution, which the actors have to accept to achieve their goal. An OPP "is viewed as the solution to a problem in terms of the resources available to the actant [actor] that proposes it as the OPP (…) It controls the resources needed to achieve the actant's outcome." [18] By defining an OPP, other possibilities are closed [16]. In the traffic example, the OPP is literally a passage, since it's the roads and the current traffic rules, etc., which have been established as a solid, reliable network.

The second moment, interessement, is where the main objective is to convince all the involved actors that the proposed problem and solution is the correct one so that they will accept to use this solution and not another one. In the traffic network, this is done by the use of sanctions from traffic rules, signs, and, not least, by the learning processes human actors go through to get a driver's license.

When the interessement of the actors is successful, the third moment, enrollment, is happening. This moment is important since it is here that support and allies are created, and the process by which actors become part of a network. The process can happen in many ways: "To describe enrollment is […] to describe the multilateral negotiations, trials of strength and tricks that accompany the interessements and enable them to succeed." [16]. In relation to the traffic network, one can think of all the things that support cars and their moving along the roads.

Finally, the last moment, mobilization, is where the actors are mobilized in such a way that they act in accordance with their prescribed roles and thereby maintain the established network. This happens when the drivers drive their cars following the rules and pathways of the traffic network.

C. *Design as inscription*

The effect of translation is delegation of power and agency. In relation to design of objects, e.g., computer games, translation is about how to construct an object in such a way that users are convinced to delegate agency. This is

described as *inscription* and *description* by Madeleine Akrich [19].

Inscription is the process where a designer embeds a special way the user has to interact with the designed object. The designer is envisaging a user and a use case for the object and develops an intended use, which is inscribed into the object by use of, for instance, physical shape, GUI, behavior of objects, and affordances in general.

Akrich compares inscription with a movie script and calls the result a script for how the user should use the object. We see thisin the design of e.g. the user interface of an iPad, where users are compelled to use finger movements to interact which are a more intuitive way of interacting and quite different from using a computer mouse.

While inscription is the designer's idea and framing of the interaction, Akrich uses the term description to describe the actual usage of the objects. This is where the script, built into and drawn upon in the design process, meets the user in an actual user setting. Coming alive is the central part of description. It is central to ANT that a non-human actor can have agency and perform actions and this is what we see when scripts, embedded in designed objects, come to life and objects engage in a network with other actors.

In the perspective of ANT, a game can be studied as a designed object with inscriptions that has agency and does something with the user, because the user invokes a network of actors and agency when he starts playing a game, i.e., following the rules of the "game world". A game designer has to be aware of the network of actors that the specific game design can invoke if he wants to be able to use it in the process of inscription. Networks of actors represent the unit of analysis in our study presented below.

III. RESEARCH METHODOLOGY

Our research method relied on qualitative data collected through observation, based on non-participatory observation as well as active participation and interviews [20], [21].We collected data from 12 game sessions during which we observed informants, recorded their behavior and interviewed them before, during, and after playing. To ensure recordable data, we used games in which players had to be social and communicate with one another and board games was particularly well suited for this since people tend to talk more when playing such games. We observed children as well as grown-ups and mixed age groups playing games in natural settings at home in a family situation or with friends.We recorded spoken language as well as body language and managed the many data using thematic and theoretical coding as described by Uwe Flick [22], who is inspired by Grounded Theory [23]. The analysis of the collected data was of course done using ANT. Researchers from social science have demonstrated that ANT is well suited for exploratory research in areas that have not been investigated much, not least because ANT-driven research is often able to draw up new conclusions [17], [24], [25]. The purpose of our study was to investigate and describe agency and actors at work when gamers play games. As our framework of analysis, we employed the concept of actors and agency and the four described moments of translation, being careful not to differentiate between non-human and human actors. We analyzed agency by following what people did with games, extracting actors and ties, and described the translation process in the actual game situations, as we will demonstrate in the next two sections. These sections are also reports of "findings" from our study. As Kraal [17] writes with reference to one of the founding fathers of ANT alongside Latour, John Law: "It is the nature of ANT that it is easier to describe through a demonstration of its use".

It is important to mention that the object of our study is not the games themselves, but the *event* that unfolds when games are played [9]. In accordance with ANT, we analyse games in action when the forces of the network are at work, so to speak.

IV. CASE: THE GAME "QUACKLE"

The case of playing the board game "Quackle" in a mixed age group is used as an example for our observations in general and in the following, we will use our analysis of this case to present our interpretation of what the game actually does.

D. Quackle! The game

The game, which was awarded "Game of the Year" in Denmark in 2006, is a typical funny board game for humans aged 5 and above. In short, the game consists of 12 different animal figures, 8 barns, and 97 playing cards with pictures of the animals and one arrow card (see Figure 1). The game starts with each player pulling an animal figure from a cloth bag showing it to the others and then hiding it in his barn so the others can no longer see it. The cards are dealt and placed in a pile in front of each player face down.

Figure 1. Photo of the game Quackle! with animals, cards, and barns on the left.

The objective of the game is to get rid of all the cards you have in your pile. Each round of the game consists of the players in turn turning a card and placing it for all to see. If two players have the same animal on their card they enter a *battle* during which the players compete on being the first to loudly say the sound of the *other* player's animal hidden in the barn. The player who looses the battle must pick his own and the pile of upwards facing cards of his opponent. The game continues until once again there are two identical animals in the cards or one of the players gets rid of all his cards [25].

The game seems pretty simple, but requires that the

players can remember and quickly mobilize the correct sounds when two identical cards are turned, which is more difficult than one might think, even for adults.

E. Game inscription

As we see in the above description of the game, there is a special way players are expected to interact with the game (the inscription) and, as we will argue in the following, in this way the game uses the learned scripts that the player brings along as well as physical and psychological abilities of the player. Among other things, the game takes advantage of the knowledge of the players (i.e., scripts) about animals and animal sounds, and the game utilizes the fact that most humans have a tendency to react automatically in pressurized situations. It is precisely this automatic reaction that makes the game funny, because the players make lots of mistakes trying to be the fastest which often result in weird sounds that is a mix between different animal sounds.

The game designer has created an inscription that can be indicated as follows: We must say a particular animal sound while we see and try to remember a lot of other animals. These many inputs are combined with the stress factor that the game introduces by stating we must respond faster than our opponents! Thus, the inscription creates a special way the player has to act, i.e., a specific way the players have to use their abilities.

In the perspective of agency, it is noteworthy that the game forces the player to make mistakes and thereby produce a mishmash of sounds which he would not normally produce. When asking our informants about the experience, most of them said their tongue was “out of control”. In this sense, it is evident that the game has agency and does something to the player.

F. Translation

The inscription plays an important role when considering the whole situation as a translation. As previously described, the translation consists of four moments which we will now outline in relation to the game scenario.

The first moment is the problematization, which is where we are presented with a problem. In our case, the game is played in natural situations on a Friday evening in a family of four (parents and two children, son aged 12 and daughter 21). For the family, the problem is the need for entertainment understood as a peaceful and enjoyable social time together. In this case, the game of Quackle is set up as a solution. Like any family game and most entertainment products, it promises that playing the game will lead to the experience of fun. Thus, the game is put forward as an actor who can do a piece of work (give us fun) through the way other actors treat it. This happens when one of the family members says, "Let's play Quackle, its fun. We always laugh so much when we play it" (quote from the daughter in this case).

The game is put forward as a solution and as the obligatory passage point (OPP) to social entertainment. The solution simultaneously suggests roles and organizes relations, i.e., a specific network where the family members will become game players and the living room table and chairs to facilitate the family sitting together. No less important is it that the game will establish equality between the players regardless of age and family position.

In the next moment, the interessement, which actually takes place in parallel with the problematization, the family members are convinced the proposed solution is the right one and barriers for alternative solutions to the problem are added. One of the things that are cut off is television; a frequently used source of entertainment in the family, when one of the adults says: "We shouldn’t watch television, we always do. We should do something together instead." (quote from the episode).

Enrollment is the third moment where the players are enrolled and this entails that they must accept the roles of participants as players of Quackle and accept the terms of the game.

In the last moment of translation, mobilization, the solution is executed when the family members sit down with the game and start playing. If the mobilization works and the translation process is thus successful, it enables the family to experience fun and laugh together. This is exactly what happened to the test family via the interaction with the game, which created a lot of laughing especially when the parents made weird sounds.

In our observations, we also encountered an event of a failed translation. In this episode, which involved four adults and two children, the setup was similar to the above but the one of the players did not accept the role of a player who could end up saying a wrong sound, and thus she ended up destroying the game. She did not hand over agency to the game and did not act as prescribed by the game.

This episode was special, but its points to the fact that the translation can fail and the players have a choice, though this choice comes with certain consequences (they never got in to play).

Going back to the situation with the successful translation, the game re-organizes the social connections within the family and in so doing builds a new network of actors and agency. The game is what Latour has named a “mediator” that “transforms, translates, distorts, and modifies” relations [12]. But the game does more than alter the social relations. It mediates the body and mind of the individual players. In the following, we will address how Quackle accomplishes the mobilization of the physical and cognitive abilities of the players.

V. WHAT THE GAME DOES

A game cannot do much by itself but is dependent on other actors, and this is, of course, particularly true for board games. Nevertheless, games have agency that makes game players act in a manner they would not have acted without the game. In that sense, the game “does” something in line with Latour’s concise statement on what defines an actor: “anything that does modify a state of affairs by making a difference is an actor [...]” [5].

Latour stresses that when we are studying a network in ANT, we are focusing on the circulation between the connections that make up the network [17]. When we look into the Quackle game, we are looking at how agency is floating between the involved actors, the details of which we

will try to demonstrate through an analysis of a play scenario.

First, the scenario of a family playing the game:

1) The game is placed on the table and the players sit down around it.

2) The game is opened, and the game elements are displayed. There are animals, barns, and cards and a cloth bag.

3) The animals are hidden in a cloth bag and all players get a barn.

4) Each player pulls an animal from the cloth bag: Player 1 gets a snake, player 2 a dog, player 3 a donkey and player 4 a frog.

5) After all animals and sounds have been introduced, they are stored out of view in the barns.

6) The cards are shuffled and dealt.

7) Everyone is ready and turn their first card.

8) A horse, a cow, a duck and a pig is turned, so there is no match.

9) Next cards are turned: a snake, a pig, a frog and an owl appears, still no match.

10) The third cards are turned: A mouse, a donkey, a rooster and an owl appear.

11) The game gathers speed and the cards are turned a bit faster.

12) The fourth card is turned: a cat, a dog, a cat and a frog.

13) Player 1 shouts "Qu..iau" [sounds a combination of a frog sound and a cat sound] and player 3 "Vu..sh"[a combination of dog sound and snake sound] followed by a grinning "Oh no, uh" and finally player 1 says "Miau" just before player 3 says "Sssshh".

14) Player 3 must gather player 1's card and the game continues.

This is the basic structure of the game which continues in a similar manner for a long time (about 30 minutes) before a player wins.

Points 1 and 2 are of practical character, but they help to create the framework for what is going to happen. Thus, the following activities are framed and the game's inscription starts to become clear, especially in the form of the rules. The agency is still with the players. This is also the case in point 3, but here the game starts to gain agency. It starts to have an effect on the players, as it prescribes their actions in the next steps.

Our observations show that, at the same time the players build up anticipation about what is going to happen which is seen by the body movements and heard by the tone and pitch of voices, this anticipation started when the players accepted the game as an OPP. It was especially noticeable in points 4 and 5, where the joy of hiding the animals in the cloth bag and pulling one provides a form of excitement that is particularly evident in the youngest child. Thus, we see here that the agency is distributed to the game as a kind of pre-disposition of body and mind [6].

In point 5, the players need to remember all the animals the other players have. The individual player has to establish links between the different animals and the players around the table. In point 7, the number of links is expanded by the creation of connection to the cards and in point 9, the game is made even more complex as more animals are introduced and it makes it harder to remember the animals hidden in the barns, which is of course part of the game designer's inscription.

We continue to point 13, where we see the first match of cards. When this match appears, a special script appears which is part of the inscription of the game. The script forces the player to act as prescribed by the game rules and it thereby functions as a type of mechanism that governs the actions of the players. The mechanism *re-organizes* the connection between the player's body and cognition in a special way by means of rules and materials (cards, animal figures, barns) and, in this manner, the game utilizes the functions of the player. As mentioned earlier, the player is driven to make mistakes when pronouncing words, and it is this "drive" that demonstrates an agency from the game.

What the game does can be described as follows: First, it mobilizes the individual player's memory but overemphasizes the need to remember. There is a wide range of images, sounds, figures, and places in play, and the player will have to revive all of these objects and connections when the match of cards happen. There are different animal figures and their sounds to choose from, and several sounds usually become actualized before the players are able to produce the correct sound.

Secondly, the game cuts across the usual connection between the player's mind and body. In point 13, it is clear that the game disrupts the usually well-controlled connections between the player's cognitive ability and their ability to control their voice. The inscription provides a procedure for a specific requested response to certain signals where the player has to use specific cognitive functions, i.e., perceive, remember, associate images and sounds as well as mobilize the organs of speech; and it all has to happen as quickly as possible. It is a simple task that players do not usually have problems with but, by adding a wide range of signals in the form of different images and sounds, and, by forcing the players to compete with others, the result is that cognitive and bodily functions respond in an incorrect manner and the players end up making wrong sounds. The game has, in a way, taken over body and mind.

The case of playing Quackle is an example of a translation process in action, where agency is delegated to a network. The case is also an example of how such a network is comprised of human, material, and social actors. The translation is only happening because the players have allowed themselves to be enrolled as players and fulfill their roles by using the material and following the rules and thereby delegating agency. In return, they are entertained.

A. Playing a computer game

Earlier in this article, we stated that we consider computer games to be games before anything else. Thus, our thesis is that computer games do something to the players when played, just as in the case of Quackle What we have

attempted until now is to establish a framework for analyzing what games do, and, in the following, we will briefly show how the framework could be applied to computer games.

The setting, which we observed, are three boys 12, 12, and 14 years old playing Grand Theft Auto V (GTA) on a Playstation 3. GTA has become very popular with its mixture of racing and adventure, where the players can follow a story already inscribed in the game or they can just go racing around in the game city.

The boys take turns at controlling the game while the two others comment and talk about what is happening. In one scenario, the 14 year old is controlling the game. He gets an assignment from the game where a tough looking guy on the screen tells him that he needs to win a race with a computer-controlled opponent to progress. Then the game begins.

The setting, we are analyzing, is a network that consists of the interior (couch, table, etc.), the Playstation (consisting of screen, game console, controller and DVD), the three boys, and the game. The game itself consists of multiple actors of which some are activated in combination with the other actors of the network.

We will not analyze all actors and possible networks the game can initiate but will only take a short look at how the game impacts the players' bodies.

When playing, the boys have to follow the rules of the game. They are complicated, but for our example here we can just point to the traffic rules in the game and how the car is driven via the controller. In the same manner as in a real traffic system, the player has to delegate agency to the system. Just as in the real traffic, there is police; here in the form of multiple cars and helicopters, and there are roads, houses, pedestrians, and the normal traffic on the road, all i of which have to be avoided during the race. All of these actors become active as the boy starts the race which lasts for a few minutes.

It is apparent how the game influences the player's body. To initiate the game, the boy presses hard on the controller and swings it forward, and the next second he and the controller are leaning heavily to the left side, almost leaning into one of the other boys. The next second, all of the boys shout "Wow, that was close!", while they all jump a little in the couch. At the end, they are all standing up and leaning forward and to the side as they follow the movements of the car on the road it tries to follow.

If we look at this scenario as a translation, we can see the problematization is set forward as the boys need to win the race and this is also the OPP. In the interessement, the game builds on the fact that the boys are already enrolled in the game (emerged in it) and thus they need to progress to keep playing. The enrollment is made more stable by the use of a character in the game and adding a storyline to the race (why they have to win), thus agency is transferred to the game. This also builds up the tension for the next moment, where the boys are mobilized to play. The term "boys" indicates that all three boys participated even though two of them did not control the game.

When the race begins, the boy controlling the game is leaning forward and swinging to the side with his body. This is where the game uses some of its agency and the bodily action of the player shows that the game is mobilizing the player's ability. In our observations, we saw this again and again, the players could not help it but move their body to the side as they turned a corner, even though in this game it was not needed, as the controller does not react to it.

The game further uses its agency when it makes the boys shout and jump. This happens as the car almost hits a wall that would have crushed the car and made them lose the game. This kind of danger is present all the time in the race. Here, the game is exercising its agency by using the player's body and mind, including his imagination that allows him and the other boys to experience danger, which in the real world would have produced fear but, in the framework of the game, produces excitement.

VI. THEORY OF PLAY AND GAMES

Obviously, excitement or pleasure is the reason why game players obey to the demands of games in the way we have described above, i.e., accept to act as a node in a network, following rules they often do not understand, using hour after hour trying to learn to manage game challenges. What games do is to produce play and playful experiences for users. In the following, we will lean on modern play theory and modern game studies to clarify the importance of play and the connection between games and play.

One need not search for long in game studies literature before it becomes evident that play, according to most researchers, is an important factor for the success of computer games as well as other kind of games. Prominent play scholars like Johan Huizinga, Roger Callois, Gadamer, and Brian Sutton-Smith appear as references in numerous articles and books on the topic. In Salen and Zimmerman's well know book on games, *Rules of Play* [27], the authors define the goal of successful game design as "...the creation of meaningful play..." [27] and later on state that "...rules are merely the means for creating play..." [27]. And to make the central point absolutely clear, they argue in a subsequent anthology on games that "...games create play: of that there is no doubt." [28]. In other words, games fulfill a function in relation to play.

In line with our view presented here is also [29], [30], and [31]. Games can be seen as "tools" that generate play, and, more importantly, games must be designed with the aim of generating play.

But what is play? In developmental psychology, play is primarily seen as a means for learning (Piaget [32], Vygotsky [33], Singer, Golinkoff, & Hirsh-Pasek [34]) and, in that frame of reference, it follows logically from the statement that games generate play that they also generate learning. Modern play theory sees play differently. Based on the work of the above-mentioned play scholars, play is regarded, in and by itself, as a meaningful human activity that we practise for the simple joy of it. Game players accept the translation of agency to games simply because they can get into play by doing so, or more accurate get into the condition in play theory called "the state of play", derived from Johan Huizinga [35] who is probably the most quoted play theoretician today. He writes in "Homo Ludens" (which translates to "man, the player") about play this way:

"…what actually is the fun of playing? Why does the baby crow with pleasure? Why does the gambler lose himself in his passion? Why is a huge crowd roused to frenzy by a football match? This intensity of, and absorption in, play finds no explanation in biological analysis. And yet in this intensity, this absorption, this power of maddening, lies the very essence, the primordial quality of play. […] … it is precisely the fun-element that characterizes the essence of play. Here we have to do with an absolutely primary category of life, familiar to everybody. […] the fun of playing resists all analysis, all logical interpretation…" [35].

The last sentence is perhaps the most important for the understanding of play and, thus, for the understanding of what games should be designed for. Play is a difficult concept to define in a scientific context because of its nature as an activity, which represents other values than the ones we traditionally use and base our thoughts on. Both in science and in our daily lives, we usually try to rationalize human activities and give them a purpose. When it comes to play, it is not possible to apply rational reasoning according to Huizinga, and play does not submit itself to the usual rational notions. We are forced to remove our accustomed patterns of thoughts and recognize that the human being is something else and more than a rational being. In short: Human beings want to play for the fun of it, and we use games primarily because they can get us "absorbed" in play.

Games, whether board games, computer games or other kind of games (of which we will present an example shortly), should be designed to facilitate this absorption. Traditional games like street games that have been around for long, some for hundreds of years, are clearly designed to produce the joy of play [31]. Games are some of the first things we meet as infants when we learn to communicate. Play researcher Brian Sutton-Smith have given a most precise definitions of play, which is useful to game design, even if it is about infants:

"[…] we postulate as the aboriginal paradigm for play, mother and infant conjoined in an expressive communicational frame within which they contrastively participate in the modulation of excitement. We call this a paradigm for all ludic action, because we suggest that other play itself is a metaphoric statement of this literal state of affairs. Ludic action, wherever it is, always involves the analogous establishment of the secure communicational frame and the manipulation of excitement arousal through contrastive actions within that frame." [36].

"Modulation of excitement" is a very precise description of what games do. There are numerous variations of such modulation. For instance, play can be physical, making the body move forward and backwards, as in sports, dancing, or on a swing; it might be psychological, creating and using a mental tension, for which jokes or horror stories are good examples. It is remarkable in this context that play if often generated by directly using the natural reactions of the body and mind, e.g., dizziness or fear, as we have tried to show in our game analysis.

We employ countless forms of materials, techniques, or genres of physical as well as immaterial types to help initiate activities that make us play. Thus, games are just one out of numerous tools [27], [29], [30] and [31]. From the simplest tools, for instance the games of dizziness, where young children turn around and around to get the excitement of dizziness, to the computer games the goal contains a familiarity. In the next section, we will present games ased on high tech, where we have utilized knowledge of games as tools for play.

VII. EXERGAMES

Exergames is one of the many names for a fairly new type of games. These games try to combine physical exercise with digital games through an interface that requires physical exertion to play the games [38], [39].

Exergames are interesting here because they combine the physical abilities of the players with the opportunities of the digital games. At the same time, many of these games are less complicated than computer games like GTA, because they rely on the physical aspects and movements of human players and less on the virtual world's narratives. This allows us to further investigate how the human players are being used within the network of a game.

In the following, we will look into one type of exergaming called modular interactive tiles ("tiles" for short).

The tiles (displayed in figure 2) are a distributed system consisting of electronic tiles, which can be assembled like puzzle pieces. The tiles combine robotics, modern artificial intelligence, and play in a product that can be used for games, sports, health promotion, rehabilitation, dance, art, and learning[39].

Every tile is 30 x 30 cm and works independently but is able to communicate with all the surrounding tiles. In this way all the tiles can communicate with each other and create a playfield for the players to play on. The tiles have a force-sensitive resistor and eight RGB light-emitting diodes able to shine in a rainbow of colors.

The many colors allow for a variety of different types of patterns and games to be played. To play a game on the tile platform, a player must move around and step on the tiles according to the rules of each game (see later). The various applications can either be played by a single person or can be set up so that multiple people can collaborate or compete against each other.

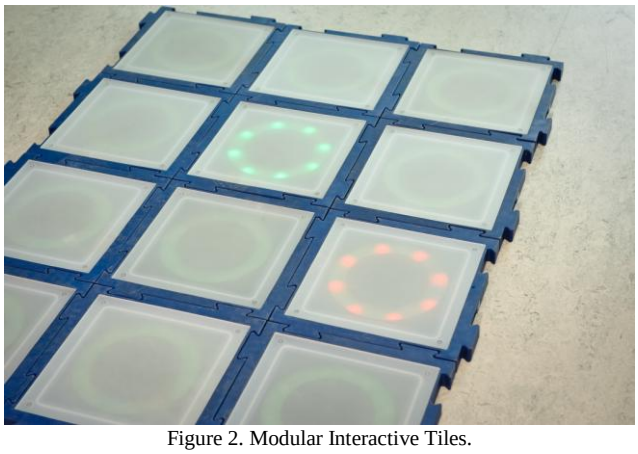


Figure 2. Modular Interactive Tiles.

Because the tiles are designed to work in any combination and because of the puzzle piece design, the tiles give the user the ability to create any playing field they wish, and to change it again anytime - e.g., change the size or shape of the field of tiles. When the user changes the configuration of the tiles, the interaction and difficulty is also changed, e.g., faster/slower movements, longer/shorter steps and so on. Thus, the user has the ability to change the movement and difficulty merely by physically building a different kind or size of the platform.

The tiles have been used as balance training for elderly people (65+ years old) and motor skills training for children (5-6 years old). We observed both elderly and children (total of 20 sessions for each group), but here we will focus on the sessions for children. Each participant participated in 10 or more sessions and a total of 19 children participated.

The data were analyzed using the methods described earlier and the following account is a prototypical example of the use of the tiles for children even though similarities exist in the use for the elderly. This example illustrates the main findings and forms a good basis for the ANT analysis. For the sake of the analysis, we are focusing on one game called "Color Race", (see [39] for more info on the tiles).

The game "Color Race" is a type of "Catch the Color" game. On the playing field, three tiles is randomly displaying different colors – red, green, and blue. Each player chose a color and has to step on the tile with the chosen color as fast possible. When they step on the tile, its color shifts randomly to another tile on the playing field that the player now has to step on.

The player has to step on tiles with their chosen color as many times as possible within a given timeframe (typically 30 sec). When the time is up, all tiles light up in the color that got most points. Hopefully, the reader can imagine three players running around on the relatively small playing field at the same time trying to step on tiles as fast as possible. The stage is set for rough-and-tumble play (in our experience regardless of age).

In the scenario that we believe is a prototypical example of the use of the tiles, we are in a kindergarten with 10 children 5-6 years old and an adult. The room is full of other toys, but there is room in the middle of the floor for the tiles. They also have chairs that some of the children are sitting on while they are waiting to play. Others are standing around and cheering or observing the children playing. The children are playing with the tiles two times a week, so they know them at this point. The adult helps to set up the tiles and they are placed in a typical setup of 9 tiles in a 3x3 square, and the game Color Race with three colors is started. Three of the children pick a color each, and they place themselves in front of that color and count down to start.

As soon as they start the game, they jump from tile to tile trying to get around the other players, but they keep bumping into each other again and again as the playing field is approx 1x1 meter so they do not have much space to move on. The game lasts for 30 seconds where the players jump around and get around 20 points each. At the end of the 30 seconds, the tiles light up in green showing that the green player got the most points.

As described above, the game requires the player to step on the tiles for the game to proceed. Here the inscription is the tiles in general and the game of "Color Race" in particular is calling for the player to step on the tiles. In our observations, we have seen this time after time. New players or observers can not resist trying to press the tiles to see what happens. The physical design of the tiles on the floor, the size of a foot, and the colorful light invite the player to step on them. They function as trigger points.

If we look at the inscription, it can be described as follows: The player must press a tile and catch as many lights as possible within a limited time frame. The game is created so the color jumps to another tile almost instantly and this creates the feeling of running after the colors, thus the name "Color Race". The movement of the light to another tile "forces" the player to act as prescribed be the game rules but also the surrounding network of competing with other children, and the observers cheering on is contributing to this "force". This is another example of what we saw earlier with Quackle!, where the players are driven to act in a certain way.

If we look at the inscription, it can be described as follows: The player must press a tile and catch as many lights as possible within a limited time frame, which organize both body and mind of each player and the interaction between them. The game also creates the necessity of speed by organizing the game as a competition. All the sessions we observed with children involved multiple players on the platform, and with more players at the same time, there is also an element of competition and a lot of communication between players. Notably, the kind of friendly communication connected to play and games. It is noteworthy that all of the players, we have observed, talk, shout and laugh. The game evokes a kind of friendly play fight.

It is of special interest from our viewpoint that the game sets up the players not only as players, but at the same time as material obstacles in the game. In the scenario with the tiles, the players are all playing at the same time and the colors jump around the platform. Here the game is using its agency. As pointed out above, the game is forcing the players to move from one tile to another, but in the process it creates a "double" role for them as players also become obstacles for other players. This "double" use of the human player is important for how the game functions. Each player becomes a game element, as they again and again are standing in the way of others who are trying to reach a tile with their color.

In the observations, we could see that exactly this point was critical for how much fun the participants seemed to get out of the game. If they surrendered to the game and accepted and maybe even used the fact that they bumped into each other, they seemed to enjoy the game more. Often players tried to push, pull or bump the other players away so they could easier reach a tile.

The game is also pushing the players to speed up and jump around by shifting the position of the light almost instantly as the tile with a color is pressed. It creates the effect of the game progressing fast, and players indicated that

they felt the need to hurry to the next tile even though the color will stay there until pressed. Technically, there was no need to hurry but mentally it appeared so.

If we consider the case as a translation we can then observe the problematization as the case of the children wanting to get into play (the state of play or play mode), and the tiles are put up as the OPP. In the interessement, the children are convinced that the tiles are the solution to the problem and the roles are divided with the children as players and obstacles for each other, the tiles as the playground and the place the game will take place.

The children and tiles are enrolled and they accept the roles in the enrollment and they accept the rules of the game, they accept that they will become both active players and obstacles in the game.

In the final moment the actual game is played. The children run around on the playing field and the tiles make them shift from one tile to another, shifting their balance, running into each other, and fighting to get the most points and by doing that clearly producing the state of play.

In this case, we tried to make it clear that the players can take multiple roles in the game, and that the actors of the network can be used both with their mental abilities (e.g., competitive revivals) as well as their physical or virtual manifestation (e.g., obstacles or trigger points).

In the following, we will go deeper into what the implications of these analyses of games in the view of ANT have for designers of games.

VIII. DESIGN IMPLICATIONS

In the introduction, we stated that games, in our point of view, could be regarded as actors because they function as organizers of other actors. Following Latour, quoted above, games are actors because they make a difference; not because they are human or non-human, social or material. We have tried to show how such "difference" is created when games do something with players. This view represents an understanding of interaction where the subject-object dichotomy is dissolved and agency is distributed in a process of reorganization, recreation and modification of actions in networks that even stretch into the mind and body of the individual player and take advantage of abilities and faculties.

If one accepts this way of viewing, it will have implications for game design, because design is not just a question of creating game worlds and interfaces but also a question of how to design social actors that can take agency and thereby initiate and guide the building of social networks, which can bring human and non-human actors to act together in such a way that the players can achieve an experience they find pleasant, joyful, funny or equivalent. As we have tried to point out, this does not only involve organizing social relations, actions and material, but also requires utilization of the player's abilities, for instance of both physical and cognitive nature.

We believe game design should be done on the basis of knowledge about how human abilities can be organized and influenced including knowledge of the abilities of different user groups. In the analysis, we showed how games orchestrate actions by humans and non-humans and resulted in experiences the players find engaging, joyful, and entertaining. From our point of view that is prototypical examples of what games do. They organize the acting of actors in order to achieve certain kinds of experiences, which, as we have argued, primarily are states of play

Through the inscription, the designer assigns agency in such a way that the game can take advantage of the characteristics of the human players. The games are examples of how the designer renders agency to a non-human object, and how these objects perform a job by getting the players to do a job.

This view gives us a possibility to further investigate how the designer can utilize this understanding when creating games.

Understanding games as active participants in the network created by or around the game, puts emphasis on attributing agency to the game and the elements in it. To understand how this is done, the concept of framing is useful.

Framing is a concept developed by Gregory Bateson [40], who points to the fact that certain situations are perceived differently than we normally would in his essay with the title "This is play" [41] which is now famous both in the context of communication and play research.

The classical example from Bateson is two monkeys playing; where in this framing a bite (an act of attacking)) does not denote what it normally would (fighting against each other) but is framed in such a way that it is percieved differently. Bateson states that a bite in the frame of play has to be followed by a metacommunicative signal "this is play", so that the opponent understands it as an act in play and not seriously meant [40], [41]. This is, for instance, the case with computer games such as GTA that we have described earlier. "This is play" puts a frame around every act which signals "not serious". But that does not mean that the acts are without influence on the players. For our viewpoint, this is a tricky point which we have to elaborate on.

The best example is perhaps the feeling of fear. Psychologist and play researcher Michael Apter [42] have put forward the example of meeting a tiger. There is a significant difference between meeting a tiger face to face in the backyard and meeting tiger in a cage, he writes in an attempt to explain that the way we experience our surroundings changes their significance due to the frame we put them into. This is especially true in play. That which outside of play would produce fear, anger and the like, does not produce the same reactions in the framework of play. Still, as the Apter example shows, what we experience in play has to *evoke* some of the same feelings as reality. If not, we would be bored. A kitten in a cage is not exciting but pittyful. We believe this is a key point in designing games. The "modulation of excitement" of course requires something to modulate. Fear is only one example. Apter writes: "One of the most interesting things about play is the tremendous variety of devices, stratagems and techniques, which people can use to obtain the pleasure of, especially to achieve high arousal […]. Putting aside the use of direct physiological interventions to increase arousal – drugs

smoking and drinking – there are a number of general psychological strategies which can be used for this purpose" [42]. A designer must know which emotions, feelings, etc. that produce arousal or other kinds of excitement and joy in the specific target group, and must know how to evoke them in a game. Good designers know that by intuition; however, explicit knowledge may help to make games better or to better avoid failures.

In terms of a game taking agency, the key point is to set the scene for the game; creating a framing where the players are willing to invest time and energy into the game and in the process distributing agency to the game. The players also have to accept the roles and rules of the game. Often this framing is done in the terms of narratives where the designer includes a story that frames the game and divides the roles.

Dividing the roles and hereby building the social network is an important part of the work done by games. This is also the first part of the translation.

We described this in the case of Quackle and how it divided special roles. This is especially clear in GTA and the case of the tiles. In GTA, the social network is built to include the actors of the race but also draw on the bigger picture of why players have to advance through the race. In the case of the tiles, the social network is constructed to create a social awareness of the actors and how they compete and play around with each other.

B. A word on scripts

The social networks and relations, actions, and materials are not the only elements to take into considerations. The most vital part that the ANT analysis points to, is to take the abilities, feelings and emotions of the players (physical as well as psychological) into account. As described earlier in the inscription, the designer can take advantage of the scripts that the players already have "downloaded", e.g., the fear of tigers, to mention a simple script.

In the example of Quackle, it was the ability to make the sounds of the animals combined with a common script that made us laugh when we and other people made mistakes inside the frame of play. In the case of the tiles, it was the game structure of "Color Race" where the players had to "catch the color" combined with the script of playful fight. Players know this kind of game; they know how it is played and the designer can use this knowledge.

All these examples are scripts in different types. As described earlier, scripts are a form of manuscripts that we know and which we use to interact and cope with different situations. In a sense, scripts can be seen as a form of recipes.

In that sense, games are dependent on the players. Players have many different scripts and understandings of how to play and what a game is. All these can be seen as part of their play culture. When players play a game or observe others playing, they learn new ways of playing and interacting: new scripts are passed to them.

It is sometimes easy to see, as when a child looks at elder children playing and starts to mimic their behavior. In this situation the child is starting to "download" the script for their actions and can later reuse these.

In all these small scripts, we have learned that the designers of game are using them in different ways while they are at the same time supplying new ones to the players.

IX. CONCLUSION AND FUTURE WORK

The main theme of this paper has been to establish an understanding of what games do in the perspective of ANT. We have seen how games do an active job and work as what Latour calls a mediator that can "transform, translate, distort, and modify" relations [12]. We believe that ANT is beneficial when we look into computer game design. While it can seem trivial that games do something to users, it is highly important for game designers to understand how games do this and why people are willing to invest time and effort in games.

We have demonstrated that, using ANT as a tool for analysis, can give us a new understanding of the interaction between games and users. We believe that game designers can advance interaction design by "following the actors" and by understanding how agency in games works, and by gaining insight into how certain bodily, psychological, and social acts can create play. We are fully aware that our analysis has shortcomings since it only covers three games although several instances of them and, thus, only a few examples of the kind of actor network which creates play. There are numerous other examples of this kind of network operating in many different ways in games.

Future work should focus on identifying, characterising, and possibly systemizing actor networks in different games. It should also focus on identifying different kinds of key scripts that the designer can utilize and take advantage of. Similarly, it's interesting to further investigate how the understanding of games as translation can help create a better awareness of what is going on in the process of game description.

ACKNOWLEDGMENT

We would like to thank our colleagues at Center for Playware, the participating children, families and elderly that allowed us to observe their play.

REFERENCES

[1] J. D. Jessen and C. Jessen, "What games do," in Proceedings of ACHI 2014, vol. 978-1-61208-325-4, pp. 222-224, 2014.

[2] P. Dourish, "Where the Action is – The foundation of Embodied Interaction. Cambrige: The MIT Press, 2004.

[3] Valve Corporation, *Counter Strike*. Washington: Valve Corporation, 2011.

[4] Blizzard Entertainment, *World of Warcraft.* Irvine: Blizzard Entertainment

[5] B. Latour, Reassembling the social: an introduction to Actor–network theory. Oxford: University Press, 2005.

[6] A. Yaneva, "Making the Social Hold: Towards an Actor-Network Theory of Design," in Design and Culture, no. 3, pp. 273-288, 2009.

[7] M. Cypher and I. Richardson, "An actor-network approach to games and virtual environments," in CyberGames '06: "Proceedings of the 2006 international conference on Game research and development," pp. 254-259, 2006.

[8] K. Kallio, F. Mäyrä, and K. Kaipainen, "At Least Nine Ways

to Play: Approaching Gamer Mentalities," Games and Culture, vol. 6, no. 4, pp. 327-353, 2011

[9] S. Giddings, "Events and Collusions A Glossary for the Microethnography of Video Game Play," Games and Culture, 4(2), pp 144-157, 2009

[10] U. Plesner, "Researching Virtual Worlds: Methodologies for Studying Emergent Practices," Routledge Studies in New Media and Cyberculture, vol. 14, 2013.

[11] M. Callon and B. Latour, "Unscrewing the Big Leviathan: how actors macrostructure reality and how sociologists help them to do so," in K. D. Knorr-Cetina and A. V. Cicourel (eds.), Advances in Social Theory and Methodology: Toward an Integration of Micro- and Macro-Sociologies. Boston: Routledge and Kegan Paul, 1981.

[12] B. Latour, "The Trouble with Actor-Network Theory," in F. Olsen, Om aktor-netværksteori. Nogle få afklaringer og mere end nogle få forviklinger." Philosophia, vol. 25 N. 3 et 4, pp. 47-64, 1996.

[13] B. Latour, "Where are the Missing Masses? The Sociology of a Few Mundane Artifacts," W. E. Bijker and J. Law (eds.), Shaping Technology/Building Society. Cambridge: The MIT Presse, 1992.

[14] B. Latour, "A Door Must be Either Open or Shut: A Little Philosophy of Techniques," in A. Feenberg and A. Hannay (eds.), The Politics of Knowledge. Bloomington: Indiana University Press, 1995

[15] O Hanseth and E. Monteiro, Understanding Information Infrastructure, University of Oslo [online]. Available from: http://heim.ifi.uio.no/oleha/Publications/bok.html 2014.01.16

[16] M. Callon, "Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay," in J. Law (eds.), Power, Action and Belief: A New Sociology of Knowledge. London:Routledge & Kegan Paul, 1986

[17] B. J. Kraal, "Actor-network inspired design research: Methodology and reflections," in Proceedings Internationl Association of Societies for Design Research, Hong Kong,, pp. 1-12, 2007.

[18] J. Rhodes, "Using Actor-Network Theory to Trace an ICT (Telecenter) Implementation Trajectory," in Information Technologies & International Development, vol 5, issue 3, pp. 1-20, 2009.

[19] M. Akrich, "The De-scription of Technical Objects," in W. Bijker and J. Law (eds.), Shaping Technology/Building Society: Studies in Sociotechnical Change. Cambridge: The MIT Presse, 1992.

[20] J. P. Spradley, Participant Observation." Orlando, Florida: Harcourt College Publishers, pp. 58-62, 1980.

[21] P. Atkinson and M. Hammersley, "Ethnography and Participant Observation," in N.K. Denzin and Y.S. Lincoln (Eds.), Handbook of Qualitative Research, pp. 248-261. Thousand Oaks: Sage Publications, 1994.

[22] U. Flick, An Introduction to Qualitative Research, 3rd edition, London,: Thousand Oaks, 2006.

[23] Juliet Corbin and Anselm L. Strauss, Basics of Qualitative Research: Grounded Theory Procedures and Techniques, 3rd edition. Sage, 2008

[24] R. Dankert, "Using Actor-Network Theory (ANT) doing research," in Publicaties vanaf, 2010 [online]. Available at http://ritskedankert.nl/publicaties/2010/-item/using-actor-network-theory-ant-doing-research 2014.01.12

[25] R. Nimmo, "Actor-network theory and methodology: social research in a more-than-humanworld," in Methodological Innovations Online 6(3), pp 108-119, 2011.

[26] Algaspel, "Quacklemanual," Algaspel, Stockholm, 2011.

[27] K. Salen and E. Zimmerman, Rules of play: game design fundamentals. Cambridge, The MIT Presse, 2004.

[28] K. Salen and E. Zimmerman, The Game Design Reader: A Rules of Play Anthology. Cambridge: MIT Press, 2005

[29] C. Jessen, Interpretive communities. The reception of computer games by children and the young, Odense University, 1999 [online]. Available at: http://www.carsten-jessen.dk/intercom.html 2014.01.12

[30] H. S. Karoff and C. Jessen, New Play Culture and Playware, in Proceedings for BIN2008, Copenhagen, 2008 [online]. Available at: http://vbn.aau.dk/files/73392625/-3BINjessenkaroff.pdf 2014.01.15

[31] H. H. Lund and C. Jessen, Playware - intelligent technology for children's play. Technical report, Mærsk Insttitute, University of Southern Denmark, 2005 [online]. Available at http://www.carsten-jessen.dk/playware-article1.pdf 2014.01.14

[32] J. Piaget, The psychology of the child. New York: Basic Books, 1972

[33] L. S. Vygotsky, "Play and Its Role in the Mental Development of the Child," in Soviet Psychology 5, pp. 6–18, 1967

[34] D. Singer, R. M. Golinkoff and K. Hirsh-Pasek (Eds.), Play=Learning: How play motivates and enhances children's cognitive and social-emotional growth. New York, NY: Oxford University Press, 2006.

[35] J. Huizinga, Homo Ludens: A Study of the Play Element in Culture. Beacon Press, Boston, 1955.

[36] B. Sutton-Smith, The Ambiguity of Play. Cambridge, Ma: Harvard University Press, 1997.

[37] H. Rodriguez, "The Playful and the Serious: An approximation to Huizinga's Homo Ludens," in Game Studies, vol. 6 iss 1, December 2006.

[38] L. H. Larsen, L. Schou, H. H. Lund and H. Langberg, "The Physical Effect of Exergames in Healthy Elderly—A Systematic Review"in Games for Health, 2(4), 2013.

[39] H. H. Lund and J. D. Jessen, "Effects of Short-Term Training of Community-Dwelling Elderly with Modular Interactive Tiles," in Games for Health, 3(5), 2014.

[40] G. Bateson, "A Theory of Play and Fantasy," in Salen, Katie og Zimmermand, Eric (eds.), The Game Design Reader: A Rules of Play Anthology. Cambridge: The MIT Presse, 2006.

[41] G. Bateson, "The message 'this is play," in B. Schaffner (Ed.), Group processes: Transactions of the second conference, pp. 145–242. New York: Josiah Macy, Jr. Foundation, 1956.

[42] M. J. Apter and J. H. Kerr, "A Structural Phenomenology of Play," in John H. Kerr and Michael J. Apter (ed.), Adult Play. Amsterdam: Swets and Zeitlinger, 1991.

Collaborative Behaviour Modelling of Virtual Agents using Communication in a Mixed Human-Agent Teamwork

Mukesh Barange, Alexandre Kabil, Camille De Keukelaere, and Pierre Chevaillier

ENIB (UEB), Lab-STICC
Brest, France
{barange, kabil, dekeukelaere, chevaillier}@enib.fr

***Abstract*—The coordination is an essential ingredient for the mixed human-agent teamwork. It requires team members to share knowledge to establish common grounding and mutual awareness among them. In this paper, we proposed a collaborative conversational belief-desire-intention (*C^2BDI*) behavioural agent architecture that allows to enhance the knowledge sharing using natural language communication between team members. We defined collaborative conversation protocols that provide proactive behaviour to agents for the coordination between team members. Furthermore, to endow the communication capabilities to *C^2BDI* agent, we described the information state based approach for the natural language processing of the utterances. We have applied the proposed architecture to a real scenario in a collaborative virtual environment for training. Our solution enables the user to coordinate with other team members.**

Keywords-Human interaction with autonomous agents, Cooperation, Dialogue Management, Decision-Making

I. INTRODUCTION

In collaborative virtual environments (VE) for training, human users, namely learners, work together with autonomous agents to perform a collective activity [1]. The educational objective is not only to learn the task, but also to acquire social skills in order to be efficient in the coordination of the activity with other team members [2]. Effective coordination improves productivity, and reduces individual and team errors. The ability to coordinate one's activity with others relies on two complementary processes: common grounding [3] and mutual awareness [4]. Common grounding leads team members to share a common point about their collective goals, plans and resources they can use to achieve them [3]. Mutual awareness means that team members act to get information about others' activities by direct perception, information seeking or through dialogues, and to provide information about theirs [4].

The collaboration in a human-agent teamwork poses many important challenges. First, there exists no global resource that human team members and virtual agents can rely on to share their knowledge, whereas, in a team of autonomous agents, the coordination can be achieved through the means of a mediator, or blackboard mechanism. Second, the structure of the coordination between human-agent team members is open by nature: virtual agents need to adopt the variability of human behaviour, as users may not necessarily strictly follow the rules of coordination. In contrast, in agent-agent interactions, agents follow the rigid structure of coordination protocols (e.g., contract net protocol). Thus, the ability to coordinate with human team members requires to reason about their shared actions, and situations where team members need the coordination to progress towards the team goal. Moreover, another important characteristic of the human-human teamwork is that the team members pro-actively provide information needed by other team members based on the anticipation of other's needs of information [5]. Thus, in a human-agent team, agents should allow human team members to adjust their autonomy and help them to progress in their task. Thus, an effective solution, supporting human-agent communications, is highly needed in a mixed human-agent teamwork. Furthermore, to exhibit the natural language communication capability, an important challenge is that the agents must take into account not only the current context of the ongoing dialogues, but also about the current context of the task and the beliefs about other team members.

This paper is the continuation and the extension of the work presented in [1]. The paper focuses on the task-oriented, collaborative conversational behaviour of virtual agents in a mixed human-agent team. Other aspects of embodied virtual agents, such as emotions, facial expressions, non-verbal communication, etc. are out of the scope of this study. As the team members must have the shared understanding of skills, goals and intentions of other team members, we proposed a belief-desire-intention based (BDI-like) agent architecture named as *Collaborative-Conversational BDI agent architecture* (C^2BDI). On the one hand, this architecture provides the deliberative behaviour for the realisation of collective activity and, on the other hand, it provides conversational behaviour for the dialogue planning to exhibit human like natural language communication behaviour for coordination. The contributions of this paper include: (1) a decision-making mechanism, in which the dialogues and the beliefs about other agents are used to guide the action selection mechanism for agents to collaborate with their team members. (2) the definition of collaborative communication protocols to establish mutual awareness and common grounding among team members; and (3) the information state based natural language processing for the task-oriented multiparty conversation. The approach consists in formalizing the conversational behaviour of the agent related to the coordination of the activity, which reduces the necessity to explicitly define communicative actions in

the action plan of the agent. It also makes the human-agent interaction more adaptive.

In Section II, we present related work on human-agent teamwork. Section III presents different components of the proposed C^{2}BDI architecture. The information state based context model is presented in Section IV. Section V describes the decision making mechanism of C^{2}BDI agent that provides the interleaving between deliberation and conversational behaviour of the agent. The collaborative conversational protocols are presented in Section VI. The natural language processing in C^{2}BDI agent is presented in Section VII. The next section illustrates how the solution fulfils the requirements of real educational scenarios. The discussion over the comparison of C^{2}BDI agent with existing approaches is presented in Section IX. Finally, Section X summarises our positioning.

II. Related Work

Both AI and dialogue literature agree upon the fact that to coordinate their activities, agents must have the joint-intention towards the group to achieve collective goal [6] and must agree upon the common plan of action [7]. Cohen and Levesque proposed the joint-intention theory, which specifies that the agents must have common intentions towards the group goal [6]. This theory does not guarantee that agents follow the same action plan. Comparing to this theory, the shared-plan theory proposed by Grosz and Kraus [7] specifies that even agents share a common action plan to achieve the group goal, it does not guarantee that agents have the commitment towards the group to achieve that goal. Both of these theories are mainly applied for the coordination among a group of artificial agents. The C^{2}BDI architecture takes the advantage of both of these theories to establish common grounding and mutual awareness among mixed human-agent team members.

A number of human-agent team models have been proposed in the literature [8]–[10]. Rich and Sidner proposed the Collagen agent [8] and Disco for Games (D4g) that is a successor of Collagen [9], which are built upon the human discourse theory and can collaborate with a user to solve domain problems, such as planning a travel itinerary, to generate dialogue about baseball and user can communicate with agents by selecting the graphical menus. In [10], Bradshaw et al. described the teamwork notification policies based collaboration model. In their model, when an important event occurs, the agent may notify the user with respect to appropriate modality and the user's position. To achieve collaboration between team members, Wooldridge and Jennings proposed a four stage model collaboration model [11] that includes (i) recognition of the potential for cooperation, (ii) team formation (iii) plan formation, and (iv) plan execution. Based on this model, Dignum et al. proposed an agent model and define how collective intentions from the team formation stage are built up from persuasion and information-seeking speech act based dialogues, using motivational attributes goal and intention [12]. Moreover, Blaylock and Allen proposed an agent based dialogue system by providing dialogue acts for collaborative problem solving to model communication at the utterance level, between a user and a system that focuses only on establishing coordination at the beginning of the shared activity [13]. Comparing to this approach, C^{2}BDI agents coordinate with team members not only at the beginning but also during the realisation of the shared task. Recently, Kamali et al. [14] have proposed a theoretical framework for proactive information exchange in agent teamwork to establish shared mental model using shared-plan approach [7].

Among many other approaches, such as speech act [15] or plan-based [8], [9], [16], the information state (IS) based approach [17] is one of the prominent approaches for dialogue modelling. It contains contextual information about the current conversation. Bunt has defined the IS, which contains contextual information of dialogue that includes dialogue, semantic, cognitive, perceptual, and social context [18], [19]. This context model includes major aspects to control natural language dialogues. However, it does not include contextual information about the shared task being carried out by the agent [20]. This leads to an incoherence between dialogue context and shared task in progress. Kopp and Pfeiffer-Lessmann proposed an IS based interaction model for *Max* agent [20]. They considered coordination as an implicit characteristic of team members, Moreover, Bunt proposed a taxonomy of dialogue acts (DIT++) based on the dialogue interpretation theory [19]. The semantics of these dialogue acts are based on the IS based approach. This taxonomy was built mainly to annotate natural language dialogues. We are motivated to use it to understand and interpret conversation between human-agent team due to its following characteristics: (i) it is mainly used for dialogue interpretation in human-human conversation; (ii) it supports task oriented conversation; and (iii) it has become the ISO 24617-2 international standard for dialogue interpretation using dialogue acts.

III. C^{2}BDI Agent Architecture

In this section, we describe components of C^{2}BDI agent architecture that provide deliberative and conversational behaviours for collaboration (Fig. 1). The C^{2}BDI agent architecture is based on the *Belief, Desire, Intention* architecture (BDI) [21] and treats both deliberative and conversational behaviours uniformly as guided by the goal-directed shared activity. The originality, compared to pure BDI, lies first on the role of dialogue, that modifies together the believes, the desire and the intentions of the agent, and second on the collaborative nature of the agent's activity. Different components of the architecture are summarised as follow:

a) Decision Making: In *C^{2}BDI* agent, the decision making includes deliberation control, and reactive behaviour control modules.

Deliberation control: Its main task is to decide how can the agent deliberate its goal to decide which one should be pursued. The decision process is driven by the information about the goals, activity plans, Information-state and the semantic knowledge of VE and of the task.

Reactive behaviour control: It uses the multimodal perception information from perception memory to reason about whether participants are in contact with, are they visible, whether someone is talking in the group, and whether the agent have a turn to talk. It implements the limited multi-model features of YMIR architecture [22] to manage multi-party conversation in particular to manage turn taking behaviour.

b) Knowledge Base: The organisation of knowledge in C^{2}BDI agent allows to establish the strong coupling be-

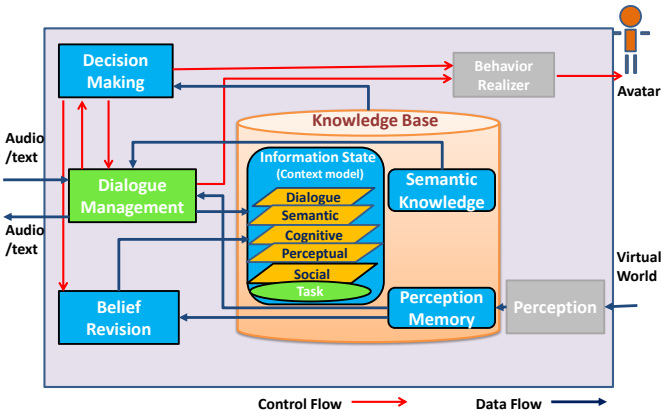


Figure 1: C^{2}BDI Agent Architecture

tween decision making and the collaborative conversational behaviour of the agent. The knowledge base consists of semantic knowledge, perception memory and IS. The semantic knowledge contains semantic information that is known a priori by the agent, such as the knowledge concerning concepts, and individual and shared plans. Following the shared-plan theory [7], C^{2}BDI agents share the same semantic knowledge about the VE and the group activity. This simplifies the planning process of agents, as agents need to construct only their local plan. Moreover, sharing the same semantic knowledge also supports proactive conversation behaviour of the agent as it allows the decision making process to identify collaborative situations and information needed by other team members. The perception memory acquires information about the state of the VE perceived by the perception module. This memory contains the belief about the state and properties of the entities in VE, and the state and actions of the team members. The IS contains contextual information about the current activity and dialogues.

c) Belief Revision: It specialises the belief revision function of BDI [21] by using the capabilities of the agent, resources used in the activity, and the Information-state. It maintains the consistency of both the knowledge base and of the Information-state by updating agent's beliefs about the current state of the world, resources and capabilities of team members using current perceptions. In the classical BDI architecture, the *belief-revision* is the internal component of the decision making module, however, in C^{2}BDI architecture, the *belief-revision* is placed outside. The reason behind this is that in C^{2}BDI agent, the beliefs are updated not only from the decisions made by the agent, but also, from the information perceived by the agent.

d) Dialogue Management: The dialogue manager allows an agent to share its knowledge with other team members using natural language communication. It supports both reactive and proactive conversation behaviours, and ensures coordination of the activity. In C^{2}BDI agent architecture, the natural language understanding (NLU) and generation (NLG) of spoken dialogues is based on the rule based approach [23]. When the agent receives an utterance, it uses NLU rules to determine the corresponding dialogue act [18], [19]. It identifies dialogue contents using semantic knowledge and contextual information from IS. The dialogue manager processes these dialogue acts and updates IS based on update rules similar to [17]. When the agent has communicative intentions, it constructs dialogue act moves and update its IS. NLG rules are used to generate natural language utterance corresponding to these dialogue moves based on the current context from IS.

e) Perception: The C^{2}BDI agent perceives VE through the perception module. The current perceived state of VE is an instantiation of concepts the agent holds in its semantic knowledge. The perception module allows agents to enrich their knowledge, and to monitor the progress of the shared activity.

f) Behavior Realiser: The behaviour realiser module is responsible for the execution of actions and the turn taking behaviour of the agent.

IV. Information State based Extended Context Model

In this section, we present the proposed context model that allows a C^{2}BDI agent to store and maintain information necessary for the decision making, and natural language conversation.

The *IS* is primarily used in literature to control natural language dialogues [17], [19]. We extended its usage as the source of knowledge between the decision-making and conversational behaviour of the C^{2}BDI agent to establish coherence between these two processes. The *IS* represents the context model of the agent, and works as an *active memory* that contains beliefs and intentions of the agent.

To participate in the task-oriented communication and to establish and maintain coordination among team members, the agent not only requires the current context of the dialogue and beliefs about the world, but also the information about the current context of the task, beliefs about other team members, and the collective attitudes. To acquire these information, we have extended the IS based context model of [19] by adding the *task context* to it (Fig.2). The extended context model includes:

- **Dialogue context:** It contains different components, that represents the features about the agents dialogue acts, and other speaker's dialogue acts. Speaker's dialogue acts contains the *utterance* received from the speaker, and the *dialogueActs* generated from the interpretation of the utterance. The agents dialogue acts contains the dialogue acts generated by the agent itself. The component *nextMoves* includes the list of dialogue moves available for the generation by the agent. Moreover, the *dialogueActHistory* stores the complete history about the agent's , and other speaker's dialogue acts, as well as about the integrated dialogue moves.
- **Semantic context:** It not only contains the agent's beliefs about the current state of the VE, but also about the current progress of the dialogues. It contains a private component that includes following features: (a) The feature *beliefs*, is instantiated from concepts

Dialogue Context	agentDialogueActs, addresseeDialogueActs, dialogueActHistory, nextMoves	
Semantic Context	agenda, proactiveAgeda, communicativePlan, beliefs, expectations	
Cognitive Context	mutual-belief	
Social Context	communication-pressure	
Perceptual Context	objectInFocus, agentInFocus, third-personInFocus, actionInFocus	
Task Context	cooperative-info	group-goal, group-desire, group-intention joint-goal, joint-desire, joint-intention, joint-commitment
	private	task-focus, goals, desires

Figure 2: Extended Information State based Context Model

the agent holds in semantic knowledge, and updated depending on the progress of the shared task. (b) The feature *agenda* contains the communicative intention of the agent. These intentions are added to the agenda due to communicative intentions generated by the realisation of the collaborative task and by the social obligations carried out by the agent. (c) The *proactiveAgenda* stores the communicative intentions of the agent generated due to the proactive communication behaviour of the agent. The agent can proactively generate the communication intention in order to establish or maintain the cooperation with other team members, to satisfy the anticipated need of informations of self or of others, or to handle the resource sharing with other team members. The semantic context also contains the information about the *expectation* that the agent can have from others. Moreover, the feature *communicativePlan* can contain a communicative plan that an agent may have to be executed.

- **Cognitive context:** It includes the mutual belief among self and other team members as the result of the mutual awareness and common grounding. The team members communicate with each other in order to establish mutual belief among them. For example, the agent establishes with other team members the mutual belief about the collective decision of the choice of the shared goal, and also for the collective decision of the plan of action to be chosen to achieve the selected goal.
- **Social context:** It includes the information about the communication pressure such as greet open, close, etc.
- **Perceptual context:** It contains information on which the agent pays attention during conversation and during the realisation of the task. The *perceptual context* contains an *attention stack*, which includes the information about the current object in focus (*objectInFocus*), actor in focus (*actorInFocus*), and also keeps the information about the third-person in focus (*thirdPersonInFocus*). In contrast to [17], the agent does not only update such information from the dialogue, but also by using information acquired in its perceptual memory. This information is particularly necessary to understand the natural language utterance in particular for the resolution of pronouns and the instantiation of contextualised semantic knowledge of the agent during NLU and NLG.
- **Task context:** It includes information about the current task in progress.

The task context is divided into two components: private, and cooperative information (cooperative-info). The *private* component of task context contains:

- *desires*, which contains the set of expected desires (expected state of the worlds) for the agent.
- *goals*, which contains a set of potential goals to be achieved individually or collectively,
- *task-focus*, which is a stack that contains the current intention of the agent about the task. The type of intentions in task-focus can any of the *Int.To*, *Int* (i.e., intention that), *Pot.Int.To* and *Pot.Int.Th* [7].

To ensure that each team member has a common intention towards the team goal, the *cooperative-info* in *task context* of IS includes beliefs about collective attitudes which includes: *group-goal*, *group-desire*, *group-intention*, *joint-goal*, *joint-desire*, *joint-intention* and *joint-commitment*. These shared mental attitudes in *task context* of an agent towards the group specifies that each member holds beliefs about the other team members, and each member mutually believes that every member has the same mental attitude. We distinguish between individual, group and joint mental attitudes of the agent.

The C^2BDI agent constructs beliefs about these mental attitudes in *collective-info* of task context in a progressive manner during the process of establishing the cooperation among team members through communication. The *group-goal* indicates that the agent knows that all team members want to achieve the goal at a time or another. Similarly, *group-desire* and *group-intention* can be defined analogously. For an agent a *group-intention* becomes a *joint-intention* when the agent knows that this intention is shared by other team members. To form a *joint-intention*, a necessary condition is that the agent must have individual intention to achieve this goal. Similarly, the semantics of joint-desire and joint-goal indicates that all team members have the same *group-desire* and *group-goal* respectively, and all team members know it. Thus, these shared mental attitudes towards the group specify that each member holds beliefs about other team members, and each member mutually believes that every member has the same mental attitude.

The *joint-intention* only ensures that each member is individually committed to acting. The agent must also ensure the commitment of others to achieve this shared goal. Agents must communicate with other team members to obtain their *joint-commitments*. The agent has a *joint-commitment* towards the group, if and only if, each member of the group has the mutual belief about the same *group-goal*, the agent has the *joint-intention* about to achieve that goal, and each agent of the group is individually committed to achieve this goal. Hence, the IS not only contains information about the current context

of the dialogue, but also that of the collaborative task, i.e., beliefs about other team members potentially useful for the agent for its decision-making.

V. Decision Making Mechanism

In C^{2}BDI agent, decision-making is governed by information about current goals, shared activity plans, and knowledge of the agent (IS and semantic knowledge). The decision making algorithm is shown in Algo. 1.

Algorithm 1 Decision making algorithm

Require: IS, *GAGT*, *GAPs*
1: B = IS.SemanticContext.Belief
2: D = IS.Task-Context.Desire
3: I = IS.Task-Context.Intention
4: agenda= IS.Semantic-context.agenda
5: proactiveAgenda= IS.Semantic-context.proactiveAgenda
6: **while** *GAGT* is not completely processed **do**
7: update-perception(ρ)
8: Compute B, D, I, and update IS
9: $\Pi \Leftarrow$ Plan(P, I)
10: **while** !$\Pi.empty()$ **do**
11: **if** agenda or proactiveAgenda is not empty or the agent has received an utterance **then**
12: Process Conversation-Behavior()
13: Compute new B, D, I , and update IS
14: $\Pi \Leftarrow$ Plan(P, I)
15: **if** the *task-focus* contains communicative intention **then**
16: Process Conversation-Behavior()
17: Compute new B, D, I , and update IS
18: $\Pi \Leftarrow$ Plan(P, I)
19: Identify-Cooperative-Situation in the current plan Π
20: **if** Cooperative-Situation is matched **then**
21: Process Conversation-behaviour()
22: $\alpha \Leftarrow$ Plan-action(Π)
23: execute(α)

The decision making process verifies whether the *agenda* in IS is not empty or if the agent has received an utterance. If so, control is passed to the conversational behaviour to that supports natural language communication. After executing the communication behaviour, the agent re-evaluates its beliefs, desire and intentions because the communication can modify the mental state of the agent through the updates in its IS.

If the *task-focus* in task context contains the communicative intention, then also the control is passed to the conversation behaviour. This situation can occur when the agent is executing some predefined conversation plan based on the current context of the task. In this case also, the agent recomputes its desire and intentions.

Otherwise, the agent chooses the plan to be realised. If It identifies cooperative situations in the collective activity where the agent cannot progress without assistance, it requires other team members to cooperate in order to achieve shared group goal. The decision making passes the control to conversation behaviour of agent in order to make establish joint commitment towards the group to achieve the goal, or when the agent needs to share the status of the goal, i.e., the goal has been achieved, or the goal is no more achievable. This situation generates communicative intentions in the *agenda* or in the *proactiveAgenda* that cause the agent to interact with team members to share their knowledge.

The agent updates its *IS* if the control is passed to the conversational behaviour, and deliberate the plan to generate a new intention. Once the intention is generated, the agent selects an action to be realised and,updates its *task-focus* in *IS* to maintain knowledge about the current context of the task.

In this procedure, it is important to note that the conversation behaviour of the agent can be called in one of the following situations:

- when the *agenda* in *semantic context* is not empty or when the agent receives an utterance from the user or from other agents. This is the reactive conversation behaviour of the agent that interprets the utterance by identifying its dialogue act (Sec. VII-B), integrates the effects of the generated dialogue act by updating different components of IS (Sec. VII-C), and generating appropriate dialogue move with respect to the speaker's dialogue act (Sec. VII-D) for the generation of natural language utterance.
- when the *proactiveAgenda* is not empty. This situation occurs in the following conditions:
  - when the agent needs the team coordination, and wants to establish group belief towards this,
  - when the agent identifies the information need of self or of others, and wants to establish group belief by providing the information or by asking for the information, respectively. For example, this situation occurs when the agent identifies the need of the resource, or wants to provide the information about resource by knowing that the addressee needs this information.
  - when the agent executes predefined conversation plan. C^{2}BDI agent exhibits the capability of executing preplanned conversation plans in the same was as the activity plan. However, one of the important difference between the conversation plan and the shared activity plan is that the conversation plan is executed locally by the host agent, and unlike shared activity plan, other team members do not monitor the progress of that plan. The agent deliberates the conversation plan and adds an intention *Int.To* to the *task-focus* in order to execute a conversation operation. The execution of the conversation operation results in updates in IS by construction of appropriate dialogue act and adding it to *agentsDialogueActs* in linguistic context, and adding corresponding communicative intention to the *proactiveAgenda*

The conversational behaviour allows a C^{2}BDI agent to share its knowledge with other team members using natural language communication, and ensures the coordination of the team activity. The agent interprets and generates the dialogues based on the semantics of dialogue acts proposed in [19] using current IS. To achieve the coordination among team members, we propose *collaborative conversational protocols* for the agent. These protocols construct the *conversational desires* for the agent which, when activated, result in *conversational*

intentions.

VI. Collaborative Conversational Protocols

As we want the agent to be proactive and cooperative, we have defined three collaborative conversational protocols (CCPs). These protocols ensure the establishment of the collaboration among team members to achieve the *group-goal*, and its end when the current goal is achieved. Every team member participating in a collaborative activity enters in the collaboration at the same time, and remains committed towards the group until the activity is finished. These protocols are modelled as the update operations in the IS based on the current context of the task and the dialogue.

A. CCP-1

When the agent has a new *group-goal* to achieve, it communicates with other team members to establish *joint-commitment*, and to ensure that every team member use the same plan to achieve the *group-goal*. Algo. 2 describes how team members collectively choose the common goal in order to establish joint-goal.

Algorithm 2 CCP1 : Collective decision for Goal Choice

Require: group G, and shared goal φ, Information state IS

—-At speaker side—-:

1: **if** Group-Intention$(G, \varphi) \wedge \neg$Mutual-belief(G, φ) **then**
2: **if** size(Group-Goals)== 1 **then**
3: $IS \Leftarrow$ addTopOfProactiveAgenda Set-Q(what-team-next-goal All)
4: **else if** size(Group-Goals) > 1 **then**
5: $IS \Leftarrow$ addTopOfProactiveAgenda Choice-Q(what-team-next-goal)
6: $IS \Leftarrow$ addExpected(team-next-goal, –, ?)
7: **else if** Receive(Inform(team-next-goal $A_j, \varphi)) \wedge$
8: Group-Intention$(G, \varphi) \wedge \neg$Mutual-belief(G, φ) **then**
9: $IS \Leftarrow$ Mutual-Belief(G, φ)
10: $IS \Leftarrow$ Joint-Goal(G, φ)
11: extract $IS \Leftarrow$ Expected(team-next-goal, –, φ)
12: $IS \Leftarrow$ addTopOfAgenda Inform(Auto-Feedback(positive-ack), All)

—-Similarly at receiver side—-:

13: **if** (Receive(Set-Q(what-team-next-goal), A_j) $\vee$
14: Receive(Choice-Q(what-team-next-goal) , A_j)) $\wedge$
15: Group-Intention$(G, \varphi) \wedge \neg$Mutual-belief (G, φ) **then**
16: $IS \Leftarrow$ addTopOfAgenda(Inform(team-next-goal $A_i, \varphi)) \wedge$
17: $IS \Leftarrow$ Mutual-Belief(G, φ)
18: $IS \Leftarrow$ Joint-Goal(G, φ)
19: **else if** Receive(what-team-next-action A_i) **then**
20: $IS \Leftarrow$ addTopOfAgenda(Inform(team-next-action(φ), A_j))

When the agent A_i has one or more *group-goals* to achieve (line 1), and if it has no mutual belief about them, it constructs *Set-Q(what-team-next-goal)* (if A_i has only one goal) or constructs *Choice-Q(what-team-next-goal)* (if A_i has more than one goal) dialogue act and addresses it to the group. This results in the addition of a communicative action to the *proactiveAgenda* in semantic context of IS. By addressing this open question, A_i allows both the user and other agents to actively participate in the conversation. If A_i receives the choice of the goal from another team member (line 7), i.e., when it receives the proposition *team-next-goal*, it adds a mutual belief about *group-goal* to its *cognitive context*, and the belief about *joint-goal* to the *task context*. It then confirms this choice by sending a positive acknowledgement (by constructing *Auto-feedback(positive-ack)*) to the speaker.

When the A_i receives *Set-Q(what-team-next-goal)* or *Choice-Q(what-team-next-goal)* from A_j, and has no mutual belief about *group-goal*, i.e., no other team member has already replied to the question (line 13), it can decide to reply to A_j based on its response time, and adds the *Inform(team-next-action)* act to $agenda$ in IS. It chooses one of its available goals from its *group-goals* of IS based on its own preference rules, and informs the team by constructing *Inform(team-next-goal)* dialogue act. When the agent receives the choice of the goal from one of the team members that matches with its potential candidate goals, it modifies its IS by adding mutual belief about *group-goal* and belief about *joint-goal*.

Now, let us consider the case when the every team member has the *joint-goal*, but no *joint-intention* towards to group to achieve *joint-goal*. Each team member can choose any of the available plans to achieve that goal. In this situation, the team members cannot monitor the activities of other team members, and thus, causes problems in establishing team coordination among them. To establish the *joint-intention* towards the group to achieve collectively chosen *joint-goal*, team members need to ensure that each team member will follow the same plan to achieve the *joint-goal*. Algo. 3 describes how team members collectively select the common plan to achieve joint-goal.

Algorithm 3 CCP1 : Collective decision for Plan choice

Require: group G, and shared goal φ, Information state IS

—-At speaker side—-:

1: **if** Joint-Goal$(G, \varphi) \wedge \neg$Joint-Intention(G, φ) **then**
2: **if** size(Plans(A_i, φ)) == 1 **then**
3: $IS \Leftarrow$ addTopOfProactiveAgenda request(Check-Q(plan-choice), All)
4: $IS \Leftarrow$ addExpected(ack, –) ▷ expectation of acknowledgement
5: **else if** size(Plans$(A_i, \varphi)) > 1$ **then**
6: $IS \Leftarrow$ addTopOfProactiveAgenda request(Choice-Q(which-plan) , All)
7: $IS \Leftarrow$ addExpected(plan-choice, –, ?)
8: **else if** receive(Inform(plan-choice A_j, $P_\varphi) \wedge$
9: Expected(plan-choice, –, ?) $\wedge$
10: Group-Intention$(G, \varphi) \wedge \neg$Mutual-belief(G, φ) **then**
11: $IS \Leftarrow$ Mutual-Belief$(G, \varphi) \wedge IS \Leftarrow$ Joint-Intention(G, φ)
12: $IS \Leftarrow$ Joint-commitment$(G, \varphi) \wedge IS \Leftarrow$ pushIntoTaskFocus(φ)
13: $IS \Leftarrow$ extract(Expected(plan-choice, –, P_φ))
14: **else if** Receive(Positive-Ack, $A_j) \wedge$ Expected(ack, –) **then**
15: $IS \Leftarrow$ Joint-Intention$(G, \varphi) \wedge IS \Leftarrow$ Joint-commitment(G, φ)
16: $IS \Leftarrow$ pushIntoTaskFocus(φ);
17: $IS \Leftarrow$ extract(Expected (ack, –))

—-Similarly at receiver side—-:

18: **if** Joint-Goal$(G, \varphi) \wedge \neg$Joint-Intention(G, φ) **then**
19: **if** Receive(Check-Q(plan-choice), A_j) **then**
20: $IS \Leftarrow$ addTopOfAgenda(Inform(confirm(plan-choice, φ),A_j)
21: $IS \Leftarrow$ Mutual-Belief(G, φ)
22: **if** Receive(Choice-Q(which-plan), A_j) **then**
23: $IS \Leftarrow$ addTopOfAgenda(Inform(plan-choice, φ),A_j)
24: $IS \Leftarrow$ Mutual-Belief(G, φ)
25: **else**
26: **if** Receive(Check-Q(plan-choice), A_j) **then**
27: $IS \Leftarrow$ addTopOfAgenda(Inform(Prefer($A_i, , \varphi, P_\varphi$):?, A_j))
28: **if** Receive(Choice-Q(which-plan), A_j) **then**
29: $IS \Leftarrow$ addTopOfAgenda(Inform(Prefer($A_i, , \varphi, P_\varphi$), A_j))

If the agent A_i has only one plan to achieve the joint-goal (line 2), it constructs *Check-Q(action-plan)* act addressing it to the group. Otherwise, if A_i has more than one plan to achieve this goal (line 4), it constructs *Choice-Q(which-plan)* act and addresses it to the group. In both the cases, A_i adds the communicative intention to the *proactiveAgenga* in *IS*. When the agent receives a choice, or the confirmation of the choice of a plan, from one of the team members, it adds *joint-intention* to its *task context*. It confirms this by sending a positive acknowledgement, and constructs the belief about *joint-commitment* towards the group to achieve *joint-goal*. When the agent receives *Choice-Q(which-plan)* or *Check-Q(action-plan)*, and has no mutual belief about *group-intention*, it constructs *Inform(plan-choice)* or *Confirm* dialogue act respectively, and adds corresponding intentions to $agenda$ in semantic context of IS to inform about its plan selection. When it receives positive acknowledgement from one of the team members, it adds individual- and joint-commitment to achieve the group-goal.

B. CCP-2

When the agent has performed all its planned actions of the shared activity, but the activity is not yet finished, the agent requests other team members to inform it when the activity will be finished. As each agent has the joint-commitment towards the group to achieve the joint-goal. That is, the team members remain committed towards the group until the goal is achieved or, the goal is unachievable. The agent can drop the goal if it believes that the goal has been achieved or no more possible. Thus, to maintain the cooperation with team members, agent can ask them to inform it if the belief about the persistent goal is modified. The protocol CCP2 is defined in Algo. 4.

Algorithm 4 CCP2

Require: Information state IS, Joint-commitment(G, φ), $P_\varphi \Leftarrow Plan(Prefer(G,\ \varphi,\ P_\varphi))$, i.e., plan preferred by G to achieve φ

—-At speaker side—-:

1: **if** Joint-commitment$(G, \varphi) \wedge$
$\exists a_x \in Plan(\varphi) \mid \{\exists a_y \in Plan(\varphi) \mid (a_x \prec a_y)\ \wedge$
$\forall a_y \in Plan(\varphi) \mid (a_x \prec a_y)\ \wedge\ Bel(A_i,\ Done(A_i, a_x), t)\ \wedge$
$\neg Able(A_i, a_y)\}$
then
2: $IS \Leftarrow$ addTopOfProactiveAgenda(directive-request
3: (Inform(goal-achieved), All)

—-Similarly, at receiver side: For all other agents $A_j \in G$ —-:

4: **if** Receive(directive-request A_i Inform(goal-achieved)) $\wedge$ Joint-commitment(G, φ)
then
5: $IS \Leftarrow$ addDesire(Inform(goal-achieved), All)

In this protocol, the agent generates *Directive-request(Inform-goal-achieved)* in its $proactiveAgenda$ to ask other members to inform it when the activity will be finished. When the agent receives this dialogue act, it adds communicative goal *Inform(goal-achieved)* to its agenda. The expression $A_i \prec A_j)$ represents that the execution of A_i is preceded by the execution of A_j. Other team members that receive this directive request, and contains the joint-commitment towards the group, modify their *IS* by adding a desire to inform about the achievement of the goal.

C. CCP-3

The agent that finished the last action of the shared activity informs other team members that the activity is terminated. The protocol CCP-3 has been described in Algo. 5.

Algorithm 5 CCP3

Require: Information state IS, Joint-commitment(G, φ), $P_\varphi \Leftarrow Plan(Prefer(G,\ \varphi,\ P_\varphi))$, i.e., plan preferred by G to achieve φ

—-At speaker side—:

1: **if** Joint-commitment$(G, \varphi) \wedge$
$\neg\exists a_y \in Plan(\varphi) \mid \forall a_x \in Plan(\varphi)(a_x \prec a_y) \wedge bel(DoneA_i, a_x)$
then
2: $IS \Leftarrow$ addTopOfProactiveAgenda(Inform(activity-finished), φ, All))
3: $IS \Leftarrow$ addBel(Group-Bel(G, Done(P_φ))

4: **if** Joint-commitment$(G, \varphi) \wedge$ Group-Bel(G, Done(P_φ) $\wedge$
desire(Inform(goal-achieved)
then
5: $IS \Leftarrow$ addTopOfProactiveAgenda(Inform(goal-achieved), φ, All))
6: $IS \Leftarrow$ addBel(Group-Bel(G, always(φ)))
7: $IS \Leftarrow$ extract(Joint-commitment(G, φ)) ;
8: $IS \Leftarrow$ extract(Joint-Goal(G, φ))
9: $IS \Leftarrow$ extract(Mutual-Belief(G, φ))

—-At speaker side: For all other agents $A_j \in G$—-:

10: **if** Receive(Inform(activity-finished) $A_j\varphi) \wedge$ Joint-commitment$(G, \varphi) \wedge$
$\exists a_y \in Plan(\varphi)\ .\ \forall a_x \in Plan(\varphi) \mid (a_x \prec a_y) \wedge\ Bel(DoneA_i, a_x)$
then
11: $IS \Leftarrow$ addBel(Group-Bel(G, $Done(P_\varphi$))

12: **if** Receive(Inform(goal-achieved) $A_i\varphi) \wedge$ Joint-commitment$(G, \varphi) \wedge$
$\neg\exists a_y \in Plan(\varphi)\ .\ \forall a_x \in Plan(\varphi) \mid (a_x \prec a_y) \wedge$
$Bel(Done,\ A_i,\ a_x)$
then
13: $IS \Leftarrow$ addBel(Group-Bel(G, always(φ)))
14: $IS \Leftarrow$ PoPTaskFocus(φ))
15: $IS \Leftarrow$ extract(Joint-commitment(G, φ));
16: $IS \Leftarrow$ extract(Joint-Goal(G, φ))
17: $IS \Leftarrow$ extract(Mutual-Belief(G, φ))

The preconditions for CCP-3 are that the agent believes that it has performed the last action of the collaborative activity, and it has the *joint-commitment* to achieve *group-goal*. If these preconditions are satisfied (line 1), it constructs *Inform(activity-finished)* dialogue act addressing it to the group, and adds this communicative intention to its *proacitveAgenda*. The predicate done(P_φ) represents that the plan P_φ has been terminated.

When the agent receives the information that the last action of the activity has been finished (line 10), and it has the belief about *joint-commitment* in its *task context*, it constructs the group belief about the status of the plan.

When an agent has group belief that the activity is finished (line 4), and has a communicative goal *Inform(goal-achieved)* to achieve (due to CCP-2), it constructs *Inform(goal-achieved)* dialogue act to inform other team members that the goal has been achieved. It then adds the belief about the achievement of the goal, and removes the corresponding intention from the *task context*. The predicate always(φ) represents that the state of the world φ remains always true, i.e., the goal φ has been achieved.

When the agent receives the information about goal achievement (line 12), it removes the corresponding intention from the *task context*, and drops the communicative goal

Inform(goal-achieved) if it has. Furthermore, it then adds the belief about the achievement of the goal, and removes the corresponding intention from the *task context*.

In a mixed human-agent team, the reaction time of each team member is different, and changes dynamically, the agent waits for certain time (until the threshold of its reaction time is expired) and if no team member has already replied, the agent can create an intention to reply. Otherwise, the agent simply listens to the conversation and updates its beliefs. Thus, in order to establish mutual awareness and to coordinate with other team members, the agent participates in the conversation. Once agents have established the *joint-commitment*, they can coordinate with other team members to achieve the *group-goal*. These protocols are instantiated when the decision-making identifies collaborative situations that satisfy necessary conditions of one of the CCPs to be fulfilled (Algo. 1, lines 19-21). These situations add expectations of information from other team members, which need to be satisfied. In a human-agent team, the user's behaviour is uncertain, i.e., a user may not necessarily follow these protocols. As the agent updates their beliefs using perception information, which can make expectations to be true from the observation of actions of user perceived by the agent, or from the information provided by other team members. This mechanism makes these protocols robust enough to deal with uncertainty about user's behaviour. One of the advantages of these protocols is that the dialogues for the coordination need not to be scripted in the definition of action plans.

VII. Natural Language Processing

The natural language processing refers to the ability to understand the natural language input utterance, integrates its meaning, and also, the ability to generate natural language utterances. The natural language understanding in C^2BDI agent includes the construction of semantic form corresponding to the input utterance (Sec. VII-A), and the interpretation of the semantic form to determine its meaning in the form of dialogue acts (Sec. VII-B). The dialogue act interpretation integrates the actual meaning of the utterance to it's IS (Sec. VII-C). The agent then selects generation rules and updates its IS in order to produce new communicative intentions (Sec. VII-D), which in tern result in generation of natural language utterances. Moreover, the C^{2}BDI agent also exhibits the capability of proactive communication (Sec. VII-E). These processes modifies the IS depends upon the role the agent plays during the conversation. In C^{2}BDI architecture, the template rule based approach is used for the natural language generation, which uses the semantic contents of the dialogue act, and the semantic information of the VE to generation natural language utterance [23].

The IS based context model is modified by the means of applying the updated rules (Fig. 3). The *Update rule* consists of a set of *precondition* and the set of effects. The *Effect* defines the possible updates on IS. All preconditions must be true to apply rules which lead to apply the updates on IS defined in effect part of rules. The *Rule-base* which contains update rules can be classified into *integrationRule* and *selectionRules*. The former is used to integrate the meaning of received utterance during dialogue act interpretation, whereas, the later is used to update the IS in order to generate natural language utterance. The *selectionRules* can be further classified into *reactive-UpdateRules* which can be applied during the generation of reactive conversation, and the *proactiveUpdateRules*, which can be applied to produce the proactive conversation in the current context of the dialogue and the shared task.

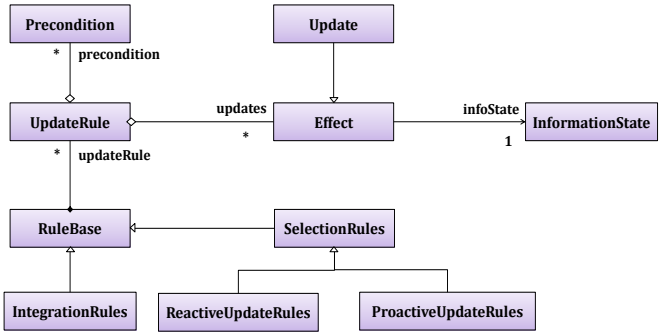


Figure 3: Information State Update Rules

In the following sections, we will describe different components of natural language processing in C^{2}BDI agent architecture.

A. Semantic form generation

The NLU focuses on the processing of the input utterance to determine its meaning. The goal is to obtain the computational form of the utterance, which can also involve the use of pragmatic aspects, and the notion of the temporality. To go further in determining the meaning, additional information is also needed to be recognised. These information or feature structure include concept types, their properties, and their relationship with other concepts in the VE, or the information about the current task. The *semanticFormGenerator* can obtain this information from the IS and semantic knowledge to generate the semantic form of utterance (Fig. 4).

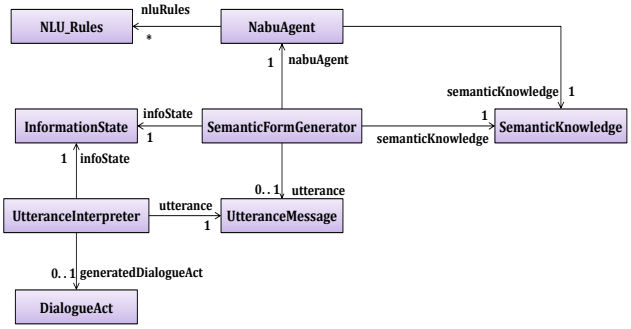


Figure 4: Utterance Interpretation

One of the important steps is the identification of the thematic roles of different components of the utterance. Identification of these roles includes information about the sender, the addressee, and the mapping of components of the utterance to the concepts in the VE, i.e., the mapping to corresponding actions, goal, concepts, entities, their features etc. In C^{2}BDI

agent, the approach is based on the template based rules (*NLU_Rules*), which are processed by *nabuAgent* (NabuTalk agent)[1]. These template based rules use the cue words, and describe the syntactic structure of the utterances. The template rule is composed of lexical expressions and parametrized functional variables, organised in appropriate order to represent the syntactical structure of the utterance. Each Lexical expression represents the regular expression to describe the cue worlds, whereas the parameterised functional variables map the components of utterances to the corresponding concepts. For example, the following simplified template rule represents the syntactical structure for the utterance of the type query.

```
(nlu-resource [id:should] #(?:sh|(?:ou?|u)l?d )#])
(nlu-resource [id:I]  [I])

(nlu-rule:
  input: {[should] [I] [concept($action)]}
 output:({[check-q] [agent-action] @my-self() @speaker()
  @concept-name($action) "next" })
)
```

In this template rule, the *nlu-resource* represents the lexical expressions to represents the lexical unit. The *nlu-rule* is composed of two components input and output. The *input* represent the template rule for the utterance, whereas the *output* represents the expression for the semantic form corresponding to the utterance to be generated. The *[concept($action)]* in *input* represents the mapping of some string to the some action, whereas, the @concept-name($ action) in *output* corresponds to the name of the action obtained through *[concept($action)]*. Now, let us consider the following sequence of dialogues: input utterance :

A_1:: ALEXANDRE: Should I place the tablet?

S_1:: SEBASTIEN: Yes, you should place the tablet.

A_2:: ALEXANDRE: Why should I do this action?

Alexandre utters A_1, addressing it to Sebastien. The input utterance is processed by Sebastien. The structure of the utterance A_1 corresponds the template input rule, the parametrized functional variable *@concept-name($ action)* is then evaluated. That is, the string *place the tablet* is mapped to one of the action or goal using the semantic knowledge.

1) Reference Resolution: Another important step towards determining the meaning of the utterance is the *reference resolution*, that is when the linguistic expression refers to the previous reference, e.g., the use of pronounces, referencing to an object or an action. The result of the *reference resolution* is that the variables that remain free are now affected to referents. The *reference resolution* requires the current context of the task and the dialogue, and the use of dialogue history. In C^{2}BDI agent, the reference of the pronoun is resolved by using information such as whether the utterance is referencing to the speaker, the addressee or to the third person. The cue-words such as *I*, *you*, and *he / she / it* are used for this purpose. For example, utterance A_1 contain the cue word *I*, thus, the receiver agent processing this utterance can identify that the speaker references herself. The agent can map the pronoun *I* to the identity of the speaker. The agent can use the contextual information stored in the *perceptual context* of IS (IV) to resolve references. The perceptual context holds information about the third person in focus, object in focus, and the action in focus during the current context of the conversation. The object resolution is done using the properties mentioned or determined by the referring expression. The action resolution refers to the action carried out by the verb or verb phrase. Solving this reference also requires the information about the current context of the ongoing activity. For the utterance A_1, the generated semantic form is shown below.

$$\underbrace{\overbrace{check-q-agent-action\ future}^{should}\quad \overbrace{@speaker()}^{I}\quad \overbrace{place\ the\ tablet}^{place-the-tablet}}_{utterance\ semantic\ form}$$

After the processing of the utterance A_1, Sebastien updates its IS and the *actionInFocus* of perceptual context now contains the action *place-the-tablet*. After uttering S_1, Sebastien processes the next received utterance A_2 that matches with the template rule given below:

```
(nlu-rule:
input: {[why] alt([should][will]) [I] [do]
alt{{[this] [action]}[this]}}
output:({[whq-why] [agent-action] @speaker()
 @concept-name($action) "future"})
)
```

Sebastien needs to resolve the action reference as the utterance A_2 contains the cue word *this action*. Since the *actionInFocus* in IS of Sebastien contains the name of the action referenced in the previous utterance, it can thus resolve the action reference by referencing it to the action stored in *actionInFocus*, which is the *place-the-tablet* for the utterance A_2.

B. Utterance interpretation

The *utteranceInterpreter* uses the semantic form of the utterance generated by *semanticFormGenerator*, and the current IS to determine the appropriate meaning of the utterance. The result of this step is the dialogue act corresponding to the utterance (Fig. 4). The agent uses the template based rules to determine the dialogue acts with reference to the semantic form of the dialogue. The dialogue act refers to the communicative function that can be understood in the dialogue context which also takes into account the previous dialogue utterances and the current context of the dialogue and ongoing activity. For example, consider a utterance:

V_1:: VIRGINIE: Yes.

According to the speech act theory, the utterance V_1 can be considered as an assertion act [15]. However, it can be precisely modelled by dialogue act as an *acknowledgement* or an *answer* act when the interpretation is associated with the previous dialogue utterance. For example, if the previous dialogue utterance is *I take the left tablet*, the utterance V_1 is the acknowledge in this case. However, if the previous utterance is *should we assemble the shelves*, the utterance V_1 is the answer to the utterance of the type *check-question*. The identification of the dialogue act requires:

- Utterance and its semantic form
- Types of the previous dialogue acts

[1]The *NabuTalk* is a commercial rule based engine that includes appropriate mechanisms to handle different NLU/NLG concepts such as utterance templates, pattern-matching, utterance understanding and generation rules

- Current context of the dialogue, and the context of the task.

The agent identifies the communicative function of the utterance, the identity of the speaker and addressee, and construct the logical form which constitute the relevant contents of the dialogue act. If the dialogue act is successfully constructed, then the utterance and the dialogue act are added to the *addresseeDialogueAct* component of the linguistic context in IS. For example, let us consider the utterance A_1 uttered by Alexandre. Sebastien processes the utterance, and identifies the associated dialogue act as an information seeking *check-question-agent-next-action* act, as Alexandre (the speaker) seeks the validity of the proposition that its next action is *place-the-tablet*. That is, the communicative function of the dialogue act is *check-question-agent-next-action*, whereas, the contents of the dialogue act includes the information about the dimension (task), speaker (Alexandre), addressee (self), and the logical form (*check-question-agent-next-action Alexandre "place the tablet"*).

C. Dialogue Act Interpretation

The result of the dialogue interpretation process is the integration of the meaning of the dialogue utterance to the context model. The formal model of dialogue act interpretation is described in Fig. 5. The dialogue act interpreter selects the update rules from the *IntegrationRules* that can be applied to the IS based on the current context of the dialogue. The dialogue act interpretation uses the current state of IS, the dialogue act, and the semantic knowledge for the evaluation of the preconditions of these rules. The successful interpretation of the dialogue act results in updates of different parts of the IS.

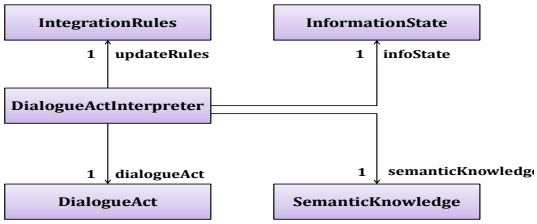


Figure 5: dialogueAct Interpretation

1) IS update when agent processes received utterance: Successful interpretation of the incoming utterance results in the processing of the DAs. Processing of the task-oriented dialogue acts provokes the changes in the *semantic context* of IS. This processing results in creating the belief about the speaker's belief, and updating the expectation of information in semantic context. The team members communicates with each others in order to establish the mutual awareness between team members. Establishing the mutual belief provokes the changes in the cognitive context. Processing of the social obligation acts will create the social pressure in social context. A successful interpretation of the utterance also results in updating the linguistic context by adding new dialogue acts to the addressee's dialogue acts. Moreover, if the utterance references to an object, addressee, sender, third person, or to the action, the perceptual context is updated. Moreover, the team members can cultivate efficient team coordination through dialogues to achieve the team goal. During this process, team members construct beliefs about different collective attitudes such as group goal, joint goal, joint commitment etc, and modifies the *cooperative-Info* component of the *task context*.

To endow C^{2}BDI agents with multiparty conversation, the updates mechanism takes into account the effects of communication on the shared mental model of team members. Consider that an agent A_i has received the utterance U_i from the speaker S_j, and $(A_i, S_i) \in G$. The utterance U_i contains the proposition P. The *UtteranceInterpretation* has identified the dialogue act D_i corresponding to the utterance U_i.

a) Processing of Information-Providing-Function: The algorithm for the context update during the processing of dialogue acts of the type *Information-Providing-Function* is given below:

1) **If** the semantic form generation or Utterance interpretation of utterance U_i is failed **Then**
   - No updates in IS.
   - Exit.
2) **If** the communicative function of DA_i is *Information-Providing-Function* **Then**
   - **If** utterance U_i is addressed to the agent A_i itself, **Then** Construct *mutual-belief* about the speaker's belief on P in Cognitive context.
     Else
   - **If** utterance U_i is addressed to the group G, **Then** Construct *group-belief* about the speaker's belief on P in *CooperativeInfo* of *task context*.
     Else
   - The receiver agent is an overhearer, thus, Construct *belief* about the speaker's belief on P in semantic context.
   - **If** utterance U_i is addressed to A_i or to the group G **Then**,
     - **If** the agent has a negative belief about P, i.e., if it believes $\neg P$, **Then** Drop $\neg P$ from semantic context.
       Else
     - **If** agent has the weak belief about P, **Then** Drop the weak belief about P from semantic context
     - Adopt the belief P, i.e. create the belief about P in semantic context.
   - **If** the agent A_i has an expectation about P from speaker, **Then**
     If the expectation about P is satisfied, **Then**
     - Drop expectation about P from *semantic context*.
     - Generate acknowledgement.
3) Copy DA_i to the dialogueActHistory in dialogue context.
4) Remove DA_i from addresseeDialogueActs in dialogue context.

b) Processing of Information-Seeking-Function: The algorithm for the context update during the processing of dialogue acts of the type *Information-Seeking-Function* is given below:

1) **If** the semantic form generation or Utterance interpretation of utterance U_i is failed **Then**

- No updates in IS.
- Exit.

2) **If** the communicative function of DA_i is *Information-Seeking-Function* **Then**
   a) **If** utterance U_i is addressed to the agent or to the group **Then**
      - Construct *mutual-belief* in *cognitive context* about the speaker's intention that the addressee provides information about P.
      - Create Pot.Int.To to reply about P to speaker.
      - Add this Pot.Int.To to the *agenda* in semantic context
      - Keep DA_i in addresseeDialogueActs

      Else
   b) The receiver agent A_i is an overhearer, therefore, Construct *belief* in *semantic context* about the speaker's intention that the addressee provides information about P.
3) Copy DA_i to the dialogueActHistory in dialogue context

D. Select and Update for Reactive Reply

At this stage, we consider that the agent has successfully interpreted the dialogue act associated with input utterance, and have updated the IS. In order to decide how to reply, the agent selects the update rules from reactiveUpdateRules that can be applied. The selection of the rules depends upon the intention in *agenda*, previous speaker's dialogue act, and the current IS.

The application of selected rules and the generation of the utterance in response to the input utterance also results in updating different components of IS. Generation of utterance with information transfer function results in updating the cognitive context or task context, depending upon whether the utterance is addressed to an addressee or to a group respectively. After the generation of utterance, the dialogue act and the generated utterance are also stored in dialogue history. After the successful processing of the intention to generate the utterance, the intention is removed from the *agenda*.

We now describe the context update algorithm when agent generate utterance in response to the incoming utterance as follows:

1) **If** Top of *agenda* is not empty
   - **If** if top of agenda contains *Pot.Int.To*
     - **If** the evaluation of conditions for *Pot.Int.To* is succeed, then upgrade *Pot.Int.To* to *Int.To*

       Else
     - pop *Pot.Int.To* from top of *agenda*
     - remove DA_i from addresseeDialogueActs in dialogue context
     - Exit.
2) Select update rules for which preconditions are true in current dialogue context and intention.
3) **If** selected rules > 0, then
   a) **ForEach** updateRule in selected rules apply update effects to IS
   b) generate and add next dialogue moves to *nextMoves* in linguistic context
   c) Pop *agenda*
   d) **ForEach** dialgoueMove in *nextMoves*
      - process dialgoueMove to generate NL utterances
      - **If** dialogueMove corresponds to the information-transfer function, then
        - **If** generated utterance is addressed to a particular addressee, then Construct the mutual belief with the addressee that the provided information is true.

          Else
        - Construct the group-belief with the group the provided information is true.

   e) Clear nextMoves;
4)
   - remove DA_i from addresseeDialogueActs in dialogue context
   - add generated dialogue to the agentDialogueActHistory

The agent evaluates the of conditions for *Pot.Int.To* before agent adopts it as *Int.To*. To do so, the agent verifies if this intention corresponding to the previous input utterance can be processed in current context of the dialogue and task. In the case when the previous utterance was addressed to the group, the agent verifies if any other agent has already replied. If so, the agent drops the intention, as the information need of the speaker has already been satisfied.

E. Proactive conversational behaviour

When the agent identifies the need of the collaboration with other team members or has identified the information need of other team members or of self, the agent can create an intention to communicate with other agents individually, or collectively, depending on the current context of the task. The agent models the proactive conversation behaviour in two steps, which are the construction of dialogue acts, and generation of next dialogue moves.

1) conversation operation: The agent executes conversation operation, which can be abstract operations such as *askOperation*, *informOperation*, *directiveRequest*, *greetOperation* etc. An extract of the conceptual model of conversation operation is shown in Fig. 6. The conversation operation can be executed if the preconditions are satisfied. The execution of the conversation operation, constructs the appropriate dialogue act, and updates the *IS* of the agent by first, adding the generated dialogue act to the *agentDialogueActs* of linguistic context, and second, it adds the associated intention Pot.Int.To to the *proactiveAgenda* in semantic context.

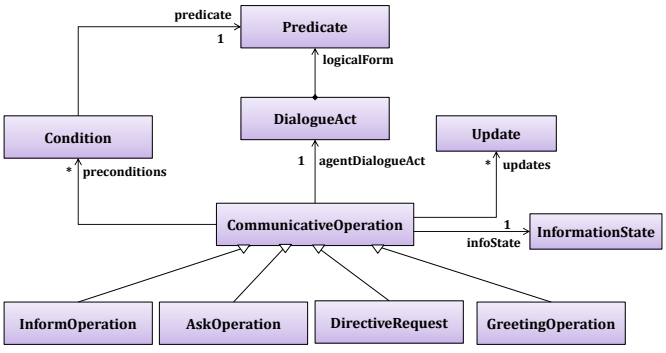


Figure 6: Conversation operation

2) IS Update for the Proactive conversational Intention : If the IS contains an intention in *proactiveAgenda*, the agent processes it. The algorithm for the context update for the proactive utterance generation is described as follows.

1) **If** top of *proactiveAgenda* is not empty, **Then If** top of proactiveAgenda contains *Pot.Int.To* , **Then**

- **If** the evaluation of conditions for *Pot.Int.To* is true, **Then** Upgrade *Pot.Int.To* to *Int.To*
  Else
- ○ Pop *Pot.Int.To* from top of *proactiveAgenda*
  ○ Remove aDA_i from *agentDialogueActs*
  ○ Exit.

2) Select update rules for which preconditions are satisfied in current dialogue context and intention
3) **If** selected (rules $>$ 0), **Then**
   - **ForEach** updateRule in selected rules
     ○ Apply update effects to IS
   - Add generated next dialogue moves to *nextMoves* in linguistic context
   - **If** communicative function of aDA_i is *Information-Seeking function*, **Then**
     ○ Aadd expectation about P from addressee
   - Pop *proactiveAgenda*
   - ForAll dialgoueMove in *nextMoves*
     ○ Process NextMove (dialogueMove, IS, aDA_i) to generate natural language utterances
   - Clear nextMoves();
4) Remove aDA_i from *agentDialogueActs* in dialogue context

The proactive conversation behaviour of the agent is driven by the information need, or by the cooperative situations where agent need to cooperate with other team members in order to achieve shared team goal. If the top of the agenda contains *Pot.Int.To*, the agent evaluates it. In this case of proactive conversation, the agent verifies if the information need of other agent or of its own, is already satisfied. If so, it drops the intention. Similarly, the agent also drops the intention to communicate if it identifies that the need of cooperation has been satisfied. Otherwise, the agent upgrades the *Pot.Int.To* to *Int.To* in order to select update rules from the *selectionRules* to update IS and to generate next moves. The agent selects rules from proactiveUpdateRules, which can be applied to *IS*, depending upon the current communicative intention, current task context, and the generated dialogue act.

The successful generation of proactive utterance addressed to an addressee (group), creates the *mutual belief* (*group-belief*) between the speaker and the addressee (group) about the speaker's information need or of addressee, depends upon the current context of the task. If the communicative function of the dialogue act is *information transfer function*, the speaker creates the mutual belief (*group-belief*) with the addressee (group) that the proposes information is true. However, if the communicative function of the dialogue act is *information seeking function*, the speaker creates an expectation of the information from the addressee (group).

VIII. IMPLEMENTATION

The technical architecture of C^{2}BDI agent is mainly composed of dialogue manager and Unity3D interface, which has been presented in [24]. Each C^{2}BDI agent is associated with a virtual human and controls its behaviors. User interacts with VE through her avatar. C^{2}BDI agent sends service messages to the associated virtual human to perform actions chosen by the decision-making module or by the dialogue manager (turn-taking behavior). The rendering system realises the requested actions and sends action events (begin, end) towards corresponding C^{2}BDI agent. The conversation manager deals with automatic-speech-recognition (ASR) and text to speech synthesis (TTS). The message manager handles the dispatching of perception information and service messages.

Figure 7: Furniture Assembly Scenario: before tablet selection

Let us now consider a motivational scenario where three agents (may include both virtual or real), named as Virginie, Sebastien, and Alexandre need to assemble a furniture. To do so, they need to choose tablets from the table (Fig. 7) and place them on shelves (Fig. 8). Following sequence of dialogues describe a typical interaction between them where a user plays the role of Alexandre.

- *S1*: Sebastien : *What should we do now?* [Set-Q(team-next-action)]
- *U1*: Alexandre : *We should place tablets on shelves.* [Inform(team-next-goal)]
- *S2*: Sebastien : *Ok.* [Auto-feedback(positive-ack)]
- *S3*: Sebastien : *Should we use the place-tablet plan?* [Check-Q(action-plan)]
- *U2*: Alexandre : *Yes.* [Auto-feedback(positive-ack)]
- *S4*: Sebastien : *I will choose the large tablet.* [Inform(resource-choice)]
  (Sebastien chooses the tablet near to him and go towards shelf;)
  (if user does not make his choice)
- *V1*: Virginie : *Alexandre which narrow tablet will you choose?* [Set-Q(what-resource-choice)]
- *U3*: Alexandre : *I will choose the left tablet.* [Inform(resource-choice)]
  (user picks the chosen tablet;)
- *V2*: Virginie : *Ok, I will choose the other one.* [Inform(resource-choice)]
  (Virginie picks the other tablet and go towards the shelf;)
  (Sebastien places his tablet on the upper position of the shelf;)
- *S5*: Sebastien : *Inform me when you will finish the activity.* [Directive-request(inform-goal-achieved)]
- *U4*: Alexandre : *Virginie which position will you use to place tablet?* [Set-Q(what-resource-choice)]
- *V3*: Virginie : *I will choose the lower position.* [Inform(resource-choice)]
  (Virginie places its tablet on the shelf)
- *U5*: Alexandre : *Ok, I will place my tablet on upper position.* [Inform(resource-choice)]
  (User places his tablet on the upper position of the shelf)
- *V4*: Virginie : *We have placed all the tablets on shelves.* [Inform(goal-achieved)]

Figure 8: Furniture Assembly Scenario right: before choosing tablet position

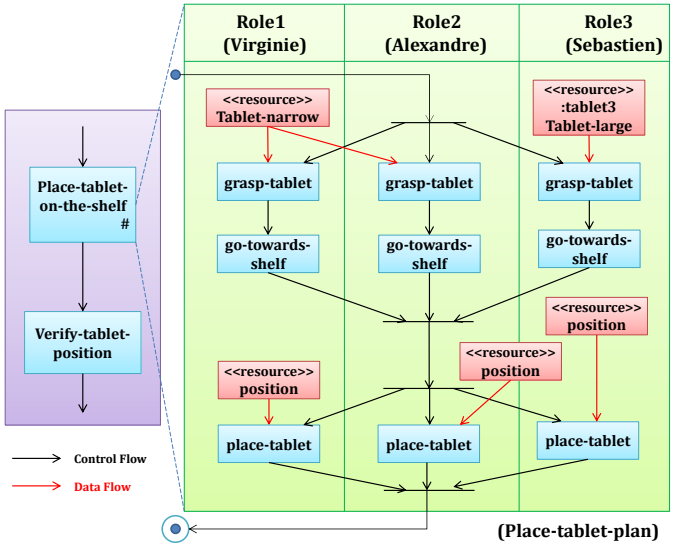


Figure 9: Partial view of Furniture Assembly plan shared between team members.

The challenging scenario includes some important characteristics such as collaborative situations to establish common grounding (*S1,U1,S2,S3,U2*), handling resource conflicts (*V1,U3,V2,*), dynamic environment (agents manipulate objects, e.g., move tablet), interleaving between communication and actions (agents utter and perform action *S4,U3,V3,U4*), mixed initiative dialogues (*V1,U3,V2* or *U4,V3,U5*), and both reactive (V3) and proactive (S1,V1) communications.

At the beginning, both user and virtual agents have a goal "place-tablet-on-the-shelf". As this goal is shared among team members, it becomes the *group-goal* (Fig. 9). A subset of knowledge of agents is shown in Table. I.

Since, Sebastien has a group-goal as *place-tablet-on-the-shelf* in its IS, but has no mutual belief about that goal, the decision making process identifies this collaborative situation

TABLE I: Snapshot of IS for Virginie and Sebastien before initialisation of CCP-1

Information State	R_1 (Virginie)	R_3 (Sebastien)
Task-Context	*cooperative-info*(group-goal ("place-tablet-on-the-shelf"))	*cooperative-info*(group-goal ("place-tablet-on-the-shelf"))

TABLE II: Snapshot of IS for agent Sebastien after establishing joint-goal

Information State	R_3 (Sebastien)
Cognitive-Context	*mutual-belief*(group-intention("place-tablet-on-the-shelf") group-goal("place-tablet-on-the-shelf"));
Task-Context	*cooperative-info*(group-goal("place-tablet-on-the-shelf") joint-goal("place-tablet-on-the-shelf"));

that fulfils conditions of CCP-1 (Algo. 1, line 19). The CCP-1 generates *Set-Q(team-next-goal)* dialogue act (Algo. 2, line 3), and adds the corresponding intention to the *agentDialogueActs*. Processing of this intention (Sec. VII-E) generates natural language utterance *S1*.

Sebastien interprets utterance *U1* as *Inform(team-next-goal "place-tablet-on-the-shelf")* dialogue act. As Sebastien has the same group-goal, it creates mutual-belief about group-goal, and generates positive acknowledgement *S2* for Alexandre. The snapshot of current state of Sebastien's IS is given in Table II. Virginie passively listens to the conversation and updates its IS following CCP-1. Now, to ensure that the each team member will follow the same action plan, Sebastien constructs *Check-Q(plan-choice)* dialogue act considering that team members have only one plan "place-tablet-plan" to achieve the current group-goal, and generates *S3*.

When both, Sebastien and Virginie receive response *U2* from Alexandre, they construct the joint-intention as well as joint-commitment towards the group-goal and update their IS. The decision making process, now, deliberate the plan and computes the new intention as *grasp-tablet* (Table III). Sebastien chooses the large-tablet as the resource is explicitly defined with the action. Virginie needs to perform explicit resource acquisition, as only the resource type is defined for its action which is dependent on Alexandre's choice (Fig. 9). As two instances of "Tablet-narrow" are available (Fig. 7), and if Virginie has no belief about Alexandre's choice, it constructs *Set-Q*(what-resource-choice) to ask Alexandre to choose one of the tablets (*V1*). When Alexandre specifies its choice (*U3*), Virginie chooses the other one (*V2*). After executing last action "place-tablet" by Sebastien from his plan, and as the shared

TABLE III: Snapshot of IS of Virginie after establishing joint-commitment

Information State	Role R_1 (Virginie)
Cognitive-context	mutual-belief(group-intention("place-tablet-on-the-shelf") group-goal("place-tablet-on-the-shelf"));
Task-Context	*cooperative-info*(group-goal("place-tablet-on-the-shelf") joint-goal("place-tablet-on-the-shelf") joint-intention("place-tablet-on-the-shelf") joint-commitment("place-tablet-on-the-shelf")); *taskFocus* (Intention("grasp-tablet") Intention("place-tablet-on-the-shelf"))

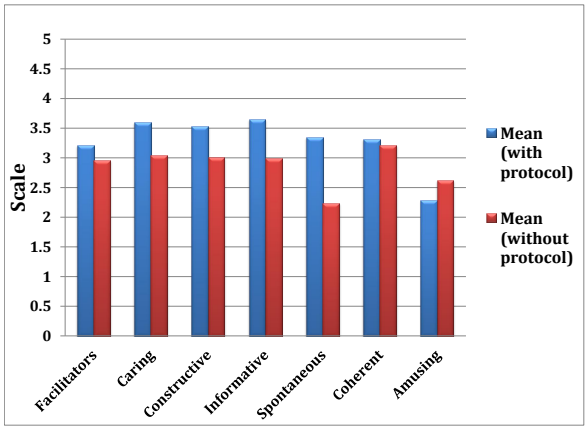


Figure 10: User evaluation: effects of communication on shared task

Figure 11: View of the collaborative scenario with one virtual team member.

activity is not yet finished, it utters *S5* following CCP-2. When Alexandre asks Virginie about its choice of position (*U4*), Virginie interprets this utterance as *Set-Q(what-resource-choice)* and informs its choice (*V3*). Once Alexandre places the tablet (*U5*) which is the last action of the shared plan, Virginie informs all the team members that the goal is achieved (*V4*) following CCP-3.

A. Evaluation

We wanted to see the contribution of conversation in a teamwork from the point of view of the user. The main aim is to see (a) the effects of conversation to establish effective team coordination, and (b) characteristics of verbal interaction with team members. We conducted the experiment in two phases. The first phase had 9 participants (group-1). Each participant was asked to perform the assembly of furniture with two virtual agents (Virgine and Sebastien) having CCPs disabled. In the second phase, 12 participants (group-2) were asked to do the same, but the virtual agents (Virgine and Sebastien) had the CCPs enabled. The participants were 3rd year engineering students between 21-23 years old. After the experiment, each participant had to respond to a questionnaire by assigning the ratings between 0 to 5 (0 means completely disagree, 5 means completely agree).

Figure 10 shows the experimental result. We found that 66% of the participants of the group-1 were agreed that the conversation with team members facilitated (mean value 2.9) them to achieve the goal. Whereas, 83% participants in the group-2 found that the conversation facilitated them (mean value 3.2) to achieve group goal. 55% participants in group-1 found that the agents took into account their participation (mean value 3), however, 75% in the group-2 found the same (mean value 3.6). The reason is that the C^{2}BDI agent takes into account the uncertainty of user behaviour, and CCPs are flexible enough to deal with this situation.

The conversation with team members was more constructive when the agents had CCPs enabled features (group-1 mean value 3, group-2 mean value 3.5). That is, virtual team members motivated them to coordinate with each other to achieve the teal goal. 55% participants in the group-1 found the communicative interaction informative (mean 2.9), whereas, 83% participants in the group-2 found it informative (mean value 3.65) as the virtual agents provided them information proactively. The reason is that the virtual agents engaged the participants in collective decision making, such as shared goal selection, plan selection, and provided the necessary information in proactive manner to perform the shared task.

Furthermore, 55% in the group-1 considered that their interaction with virtual agents was not spontaneous (mean 2.22), whereas, 75% in the group-2 found it more spontaneous (mean 3.34), that is 22.4% of gain value for spontaneous interaction. They found that virtual team members initiated the conversation driven by the information needs of participants. Moreover, both the groups found the conversation with the agents coherent (mean value 3.2, and 3.3 in group-1 and group-2 respectively) to the task. However, participants in the group-2 indicated that sometimes the conversational interaction was not amusing (mean value 2.28). The reason for this less amusement was that the participants experienced the similar conversational interaction at the time of the initialisation of each new group-goal. Nevertheless, participants in group-2 admitted that the agents' cooperative conversational behaviour helps them to effectively achieve the shared team goal.

B. Integration with Virtual Agent

The C^{2}BDI architecture has been integrated with the interaction model for virtual and real human [25] on the GVT platform [26] for learning of a procedure for the industrial maintenance [27]. This scenario describes a maintenance procedure in a plastics manufacturing workshop. The scenario consists in the replacement of a mould in a plastic injection moulding machine (Fig. 11). This specific intervention requires a precise coordination of tasks between two workers: the setter and the machine operator. The use of autonomous agents allows the learner to execute the learning procedure. The user interacts with VE by controlling his avatar thanks to a tracking system of the body and hands (Fig. 12).

Figure 12: View of the collaborative scenario with one user.

IX. Discussion

The proposed work is done based on the theoretical framework of joint intention, shared plan, and collaborative problem solving approach. These approaches aimed to specify the mental states (belied, goal, intention) during the collaboration, whereas our approach focused on the practical use of natural language dialogues for cooperation in human-agent teamwork. Moreover, these models do not specify how their model looks like. In contrast, we described an extended Information State based context model. Our belief that the team members require the belief about other members in order to establish collective intention towards the group to achieve shared team goal is close to the theoretical framework of Dignum and Dunin [28], and Dunin and Verbrugge [29] for the teamwork in multi-agent systems. In their approach, an initiator agent identifies the potential for collaboration of each team members and tries to form a team by asking confirmations from other team members, and thus follows the master slate mechanism. However, in our approach, each team member participates in collective decisions (such as the choice of a group-goal, the choice of the shared plan to achieve that goal). Moreover, team members also provide opportunities and motivations for other team members (including the user) to participate in the natural language conversion in order to establish efficient coordination among them.

The context model of C^{2}BDI agent is inspired by the context models proposed in Traum and Larsson [17], Keizer and Morante [30], and Bunt [19]. However, it has significant differences with their context model. The context models in [30] and [19] include the system belief and user's belief in semantic context and in cognitive context respectively. However, in C^{2}BDI agent, the semantic context contains the beliefs about the agent's own beliefs, and the beliefs about other team members. Moreover, the task-context in C^{2}BDI agent contains collective attitudes in the cooperative information (cooperative-info), which includes information necessary to establish and maintain coordination with other team members. Furthermore, the context models presented in [17] and [30] only accommodate an agenda that holds the communicative intentions of the agent. However, in the context model of C^{2}BDI agent, the semantic-context contains agenda and proactiveAgenda to store the intentions generated due to reactive and proactive conversation behaviour respectively. Moreover, the C^{2}BDI agent also manages the intentions to perform actions in task-focus in the task-context explicitly. Thus, the IS not only contains the current context of the dialogue but also the ongoing task of the agent.

Most of the dialogue system support two party conversation, however, the conversational behaviour of C^{2}BDI agent deals with multiparty conversation as the agent can play different roles (i.e., speaker, addressee, or overhearer) during the conversation. The information state is mainly used in these approaches to handle the conversation, and can be updated during dialogue processing. In contrast, in C^{2}BDI architecture, the information state is updated during the dialogue processing, but also during the deliberation of the task. Comparing with the context model for *Max* agent proposed by Kopp and Pfeiffer-Lessmann [20], in which the cooperation is considered as an implicit characteristic of agents, C^{2}BDI agents exhibit both reactive and proactive conversational behaviours, and explicitly handle cooperative situations through natural language communication between team members taking into account the user in the loop.

The proposed behavioural architecture can be improved in many ways. For the simplicity, we considered that an utterance contains only one communicative function. However, dialogue utterances often have multiple communicative functions, such as answering a question but also providing feedback on the understanding of the question, and also taking the turn [19]. Like Bunt [31], we are convinced that the taking into account both of these features require to define more precise update semantics for dialogue acts. Furthermore, the C^{2}BDI agent architecture does not take into account different modalities of interaction, such as facial expressions, emotions, gesture, gaze. However, these are linked in language production and perception, with their interaction contributing to felicitous communication [32]. It will be interesting to integrate these modalities in order to improve believability, usability, and coverage of interaction in a mixed human-agent teamwork.

X. Conclusion

The proposed behavioural architecture C^{2}BDI endows the agents in the collaborative VE with the ability to coordinate their activities using natural language communication. This capability allows users and agents to share their knowledge with their team members. The architecture ensures the knowledge sharing between team members by considering the deliberative and the conversation behaviours, not in isolation, but as tightly coupled components, which is a necessary condition for common grounding and mutual awareness to occur. The collaborative conversational protocols we proposed enable agents to exhibit human-like proactive conversational behaviour that helps users to participate in the collaborative activity. We proposed the information state based approach for natural language processing, in which the semantic information about VE and the shared plans is used as knowledge source. Moreover, we described the context update mechanisms to integrate the effects of both reactive and proactive conversation, bases on the role played by the team members during conversation. Furthermore, user experience also confirms the advantages of collaborative conversational behaviour of agents for the efficient team coordination in human-agent teamwork. While the implemented scenario already shows the benefits of

the solution, the behaviour of the agents could be enriched both in terms of collaborative team management and in terms of natural language dialogue modelling. Particularly, it would be interesting to endow agents with problem solving capabilities to select their communicative intentions, or to engage themselves into information seeking behaviours and negotiation rounds, as observed in human teamwork [33].

ACKNOWLEDGMENT

This work was partly supported by the ANR (Corvette project ANR-10-CORD-012).

REFERENCES

[1] M. Barange, A. Kabil, C. De Keukelaere, and P. Chevaillier, "Communicative capabilities of agents for the collaboration in a human-agent team," in *Proceedings of 7th International Conference on Advances in Computer-Human Interactions ACHI'14*, 2014, pp. 389–394.

[2] C. Barot, D. Lourdeaux, J.-M. Burkhardt, K. Amokrane, and D. Lenne, "V3S: A virtual environment for risk-management training based on human-activity models," *Presence*, vol. 22, no. 1, pp. 1–19, 2013.

[3] H. H. Clark and E. F. Schaefer, "Contributing to discourse," *Cognitive Science*, vol. 13, pp. 259–294, 1989.

[4] K. Schmidt, "The problem with 'awareness': Introductory remarks on awareness in CSCW," *Computer Supported Cooperative Work*, vol. 11, no. 3, pp. 285–298, 2002.

[5] X. Fan, J. Yen, and R. A. Volz, "A theoretical framework on proactive information exchange in agent teamwork," *Artificial Intelligence*, vol. 169, no. 1, pp. 23–97, Nov. 2005.

[6] P. R. Cohen and H. J. Levesque, "Confirmations and joint action," in *Proceedings of IJCAI'91*, 1991, pp. pages 951–957.

[7] B. J. Grosz and S. Kraus, "Collaborative plans for complex group action," *Artificial Intelligence*, vol. 86, no. 2, pp. 269 – 357, 1996.

[8] C. Rich, C. L. Sidner, and N. Lesh, "Collagen: applying collaborative discourse theory to human-computer interaction," *AI Mag.*, vol. 22, no. 4, pp. 15–25, Oct. 2001.

[9] C. Rich and C. L. Sidner, "Using collaborative discourse theory to partially automate dialogue tree authoring," in *Intelligent Virtual Agents*, ser. LNCS, Y. Nakano, M. Neff, A. Paiva, and M. Walker, Eds. Springer Berlin Heidelberg, 2012, vol. 7502, pp. 327–340.

[10] J. Bradshaw, P. Feltovich, M. Johnson, L. Bunch, M. Breedy, T. Eskridge, H. Jung, J. Lott, and A. Uszok, "Coordination in human-agent-robot teamwork," in *Collaborative Technologies and Systems, 2008. CTS 2008. International Symposium on*, 2008, pp. 467–476.

[11] M. Wooldridge and N. R. Jennings, "The cooperative problem-solving process," *J. of Logic and Computation*, vol. 9, no. 4, pp. 563–592, 1999.

[12] F. Dignum, Dunin-Keplicz, and R. Vebrugge, "Agent theory for team formation by dialogue," in *Intelligent Agents VII Agent Theories Architectures and Languages*, ser. LNCS. Springer Berlin, 2001.

[13] N. Blaylock and J. Allen, "A collaborative problem-solving model of dialogue," in *In Proceedings of the SIGdial Workshop on Discourse and Dialog*, 2005, pp. 200–211.

[14] K. Kamali, X. Fan, and J. Yen, "Towards a theory for multiparty proactive communication in agent teams." *Int. J. Cooperative Inf. Syst.*, vol. 16, no. 2, pp. 271–298, 2007.

[15] J. R. Searle, *A taxonomy of illocutionary acts*, K. Gunderson, Ed. Minneapolis: University of Minnesota Press, 1975.

[16] P. Cohen and C. R. Perrault, *Elements of a plan-based theory of speech acts*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1986, pp. 423–440.

[17] D. Traum and S. Larsson, "The information state approach to dialogue management," in *Current and New Directions in Discourse and Dialogue*, ser. Text, Speech and Language Technology, J. Kuppevelt and R. Smith, Eds. Springer Netherlands, 2003, vol. 22, pp. 325–353.

[18] H. Bunt and Y. Girard, "Designing an open, multidimensional dialogue act taxonomy," in *Proceedings of DIALOR'05*, Nancy, 2005, pp. 37–44.

[19] H. Bunt, "The semantics of dialogue acts," in *Proc. of the 9th Int. Conf. on Computational Semantics*, ser. IWCS '11, Stroudsburg, PA, USA, 2011, pp. 1–13.

[20] S. Kopp and N. Pfeiffer-Lessmann, "Functions of speaking and acting: An interaction model for collaborative construction tasks," in *D. Heylen, S. Kopp, S. Marsella, C. Pelachaud et H. Vilhjálmsson, editeurs, The First FML workshop, AAMAS*, vol. 8, Portugal, 2008.

[21] A. S. Rao and M. P. Georgeff, "Bdi agents: From theory to practice," in *1st international conference on multi-agent systems*, 1995, pp. 312–319.

[22] K. R. Thórisson, "A mind model for multimodal communicative creatures & humanoids," *International Journal of Applied Artificial Intelligence*, pp. 519–538, 1999.

[23] M. Barange, P. D. Loor, V. Louis, R. Querrec, J. Soler, T.-H. Trinh, E. Maisel, and P. Chevaillier, "Get involved in an interactive virtual tour of brest harbour: Follow the guide and participate," in *Proceedings IVA'11*, ser. LNCS, vol. 6895. Springer, 2011, pp. 93–99.

[24] M. Barange, A. Kabil, and P. Chevaillier, "The c2bdi agent architecture for teamwork coordination using spoken dialogues between virtual agents and users," in *Advances in Practical Applications of Heterogeneous Multi-Agent Systems. The PAAMS Collection*. Springer, 2014, pp. 315–318.

[25] A. Saraos Luna, V. Gouranton, and B. Arnaldi, "Collaborative Virtual Environments For Training: A Unified Interaction Model For Real Humans And Virtual Humans," in *Learning by Playing. Game-based Education System Design and Development*, 2012, pp. 1–12.

[26] S. Gerbaud, N. Mollet, F. Ganier, B. Arnaldi, and J. Tisseau, "GVT: a platform to create virtual environments for procedural training," in *IEEE Virtual Reality*, Reno Etats-Unis, 2008, pp. 225–232.

[27] T. Lopez, P. Chevaillier, V. Gouranton, P. Evrard, F. Nouviale, M. Barange, R. Bouville Berthelot, and B. Arnaldi, "Collaborative Virtual Training with Physical and Communicative Autonomous Agents," *Computer Animation and Virtual Worlds*, vol. 25, pp. 485–493, May 2014.

[28] F. Dignum, B. Dunin-Keplicz, and R. Verbrugge, "Creating collective intention through dialogue." *Logic Journal of the IGPL*, vol. 9, no. 2, pp. 289–304, 2001.

[29] B. Dunin-Keplicz and R. Verbrugge, *Teamwork in Multi-Agent Systems*. John Wiley & Sons, Ltd, 2010, ch. Dialogue in Teamwork, pp. 139–168.

[30] S. Keizer and R. Morante, "Dialogue acts as context operators constraining the generation of dialogical discourse," in *Proceedings of the Workshop on Constraints in Discourse*. Citeseer, 2006, pp. 117–124.

[31] H. Bunt, "A context-change semantics for dialogue acts," in *Computing Meaning*, ser. Text, Speech and Language Technology, H. Bunt, J. Bos, and S. Pulman, Eds. Springer Netherlands, 2014, vol. 47, pp. 177–201.

[32] P. Wagner, Z. Malisz, and S. Kopp, "Gesture and speech in interaction: An overview," *Speech Communication*, vol. 57, pp. 209–232, 2014.

[33] W. J. Clancey, "Simulating activities: relating motives, deliberation, and attentive coordination," *Cognitive Systems Research*, vol. 3, pp. 471–499, 2002.

Exploring the Digital Music Instrument Trombosonic with Extreme Users and at a Participatory Performance

Oliver Hödl and Geraldine Fitzpatrick
Human Computer Interaction Group
Institute for Design and Assessment of Technology
Vienna University of Technology
oliver@igw.tuwien.ac.at, geraldine.fitzpatrick@tuwien.ac.at

Simon Holland
Music Computing Lab
Open University
Milton Keynes, England
simon.holland@open.ac.uk

***Abstract*—We introduce the "Trombosonic" as a new digital music instrument inspired by the slide trombone. An ultrasonic sensor combined with a red laser allows the performer to play the instrument using similar movements to playing a trombone to change the pitch, despite the absence of a physical slider, by moving one hand back and forth. Additional sensors enhance the potential for musical expression by movement of the whole interface and by using the breath. We identify and discuss a variety of design issues arising from the Trombosonic. Due to its compact size and the lack of a slider, the Trombosonic can be played in many different ways. In order to explore varied potential uses of the Trombosonic, we carried out a series of informal evaluations. These included experts in new musical instruments, an older user, a younger user, an interaction design expert, and the audience at an experimental concert with audience participation. Future work is also discussed. Further technical development might include a built-in microphone to use the human voice and an expansion of the synthesizer's features.**

***Keywords*-Sound and Music Computing, Interface for Musical Expression, Digital Music Instrument, Exploratory Evaluation, Performance Evaluation**

I. INTRODUCTION

In this article we build on a recent study about the new digital music instrument "Trombosonic" [1]. New musical instruments, such as ours, are often equipped with sensor-technologies to allow many different ways of expression and interaction [2]. Apart from using them for musical purposes the application of such versatile interfaces can be manifold (e.g., [3], [4]) but remain as yet largely unexplored.

The main contribution of this paper is the presentation of the new digital music instrument Trombosonic and to discuss its potential uses, as derived from analysis of an exploratory evaluation as well as its use in a participatory performance.

The primary intention was a new digital music instrument inspired by the slide trombone. Hence, we started to design the interface under some self-defined constraints. Unlike many existing approaches, we did not augment a trombone (e.g., [5]) or used the instrument as an example for a digital music interface imitating the trombone's look and feel to create an electronic slide trombone (e.g., [6]). For our development we rather took the technique for playing the trombone as a guiding principle only, to enable an embodied control of sound with a preferably simple and compact hand-held interface.

Our initial design considerations led to preconditions that address sensor-technology and construction. To balance functionality and complexity of the interface and keep it as simple and cheap as possible, we decided to use only standard off-the-shelf low-cost hardware, such as an ultrasonic sensor, push buttons and an accelerometer, to mention some of the important ones. By doing so, we could keep the costs for hardware and material below 100 Euros in total and were still able to explore a range of different sensors within one device.

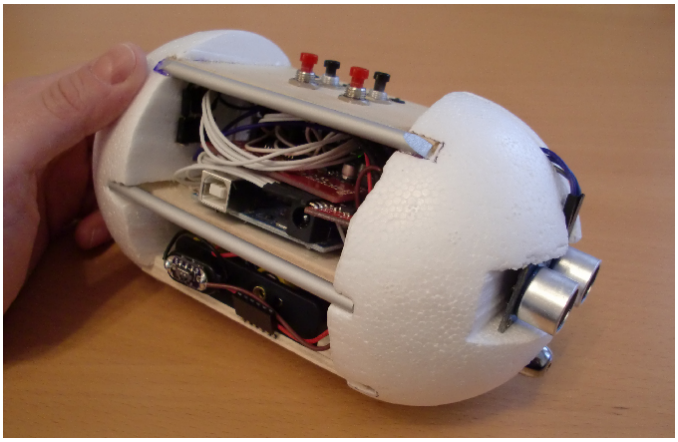


Figure 1: The final Trombosonic prototype

Throughout the design process, the basic intentions regarding appearance, functionality and materials changed significantly. We initially started with a paper-made tubular prototype to simulate a trombone. The final interface is shown in Fig. 1 and illustrates the visual difference to a traditional slide trombone (see Fig. 2). Most notable is the missing typical slide that characterises a trombone. Despite that, it is played like a trombone with slide motions by holding it in one hand, either left or right, and changing the pitch by moving the other hand back and forth. The name "Trombosonic" is the combination of the two words "trombone" and "sonic".

However, during the development phase it turned out that the device might also be useful for other applications as a hand-held interface. Apart from its original purpose to serve as a musical instrument, an exploratory evaluation has shown its potential applicability in fields such as education, sonification, therapeutic prevention and rehabilitation. We use both, expert

Figure 2: The jazz trombonist Roman Sladek plays a traditional slide trombone (Photographer OhWeh)

Figure 3: Playing the Trombosonic, the red laser in the palm indicates the direction of the ultrasonic sensor

knowledge and the concept of using the experience of extreme users [7] to identify potential future applicability in music and non-music domains.

Furthermore, we conducted an evaluation during a live performance of the first author's band "Oliver Linus". The whole performance took place at a music festival in Vienna, Austria and was planned as a participatory performance for another study where the audience was included at certain parts of the show. The actual focus of that, from a musical perspective, is discussed in another publication [8] and not directly relevant or of interest here.

In the following, we start with the description of similar research and existing literature our project is built on. Then we go on to describe the design and the functionality of the Trombosonic. Finally, we present the exploratory evaluation that shows the potential applicability of our prototype.

II. Related work

In this paper we consider design issues, an exploratory evaluation, and potential wider classes of use. Consequently, the following brief review of related work considers work related to all three aspects of this paper.

Both researchers and artists have used the trombone for their work in many ways. Composers appreciate the trombone as "very adaptable system for capturing, suspending and altering shards of sound" using different electronic extensions to create new sounds and sample external sources [9]. Farell augmented a trombone by using a minimal-hardware ultrasonic sensor for the slide, a modified mouthpiece and a loudspeaker in the trombone bell to change the original sound of the trombone for his electro-acoustic performances [5]. A very simple prototype using an optical sensor to detect the position of the slide was created by Lemouton et al. to realise a gestural interface for a traditional trombone [10].

Instead of augmenting existing instruments, Bromwich built a completely new instrument, the Metabone, using only the trombone's dynamics and characteristics [11]. Su et al. present an electronic trombone for the entertainment of children and a playful introduction in musical instruments [12]. The Double Slide Controller derives from the traditional trombone [6]. It looks different though and appears as a complex interface.

Unlike the presented examples that use the trombone as a model or augment an existing instrument, we wanted to combine its most promising features within one simple and compact interface. Keeping in mind the trombone as original instrument and its possibility to create sound by a unique hand gesture, we also wanted to provide new features and embodied interaction that goes beyond the usual musical purposes.

Apart from designing, building and playing new musical instruments, their evaluation can shed light on improvement possibilities and the experience of musicians and audiences with these new developments. Especially in Human-Computer Interaction, researchers have tried different approaches to evaluate digital music instruments, such as Kiefer et al. [13] and Stowell et al. [14].

The evaluation of new musical instruments at an actual live performance, allows researchers to access the original opinion and experience of an audience. Usually, this consists of a certain number of people and the performances happen in an authentic, real-world setting (e.g., [15]).

Approaches to get feedback from the audience are manifold, for instance, by using technology (e.g., [16]) or ethnographic methods such as questionnaires (e.g., [17]). Research motivations for gathering audience feedback are not limited to purposes of evaluating new digital instruments. The use of audience feedback is also a key technique in technologically-mediated participatory performance.

In both cases, researchers have used new technology to measure emotional states of the spectators while new digital instruments were played live in front of them (e.g., [18]). Others focused on traditional forms of feedback such as measuring applause (e.g., [19]).

Beyond pure artistic and musical purposes, digital instruments have been approached from different angles. For instance, Robson [4] and Jordà [20] have shown the suitability of certain digital music instruments as playful, toy-like devices for non-specialists. Others investigated their applicability for therapeutic prevention and rehabilitation (e.g., [21], [22]). In the context of user-driven innovation, Holmquist [7] explored the value of extreme users in design evaluation.

All these approaches from pure digital instrument design, towards evaluation in a musical context such as participatory performances, and finally studies with non-expert users in other domains than music, form the basis of the studies we present

TABLE I: Overview of prototype development and evaluation phases

Phase	Description
Development	Interface Design (Section IV)
Development	Musical Expression (Section V)
Evaluation	Exploratory Evaluation (Section VI)
Evaluation	Evaluation at a Participatory Performance (Section VII)

in this article.

We proceed with the description of the design process and considerations that have been engaged with during the prototype development. Along with that, technical details about the Trombosonic are presented from an engineering perspective as well as from the perspective of sound creation.

III. Research Approach

This research was driven by the idea to develop an interactive interface for gestural control of sound. In particular, the interface should be a digital music instrument to explore different application possibilities in music and non-music domains.

We already had certain design considerations in mind at the beginning. Hence, we decided to follow a design-led research approach to develop the interface (the first author is both a musician and interaction designer). In the first step, we focused on the overall concept and the technical development of the interface (Section IV). The creation of sound capabilities came next to turn the interface into an actual instrument (Section V).

To explore the different application possibilities of the final prototype, the "Trombosonic" was evaluated in two consecutive phases. First, we did an exploratory evaluation with extreme users (Section VI); this was followed by improvements to the synthesizer based on insights from this user study. The second evaluation was conducted at a live concert to reflect on the audience's experience (Section VII). Thus, we see this as a provocative prototype to help us explore a design space of possibilities for this type of new digital music instrument enabled by sensor-based technologies.

An overview of all phases including the prototype development and the evaluation is presented in Table I. All phases are described in detail in the following sections.

IV. Interface Design

A. Design Process

Throughout the design process, the basic intentions regarding appearance, functionality and materials changed significantly. We initially started with a paper-made tubular prototype to simulate a trombone. The original setting is documented in Fig. 4. To create a trombone-like hand-held interfaces, we used two interleaved paper tubes with an ultrasonic sensor at one end. The other end was left open similar to a mouth piece as known from wind instruments.

Right from the beginning we used an Arduino Duemilanove microprocessor [23] for sensor handling. To connect our prototype to a MIDI compliant synthesizer for sound testing purposes, we implemented a simple MIDI interface on the Arduino. Later, this wired MIDI interface was replaced by a wireless OSC interface to increase physical and technical operability.

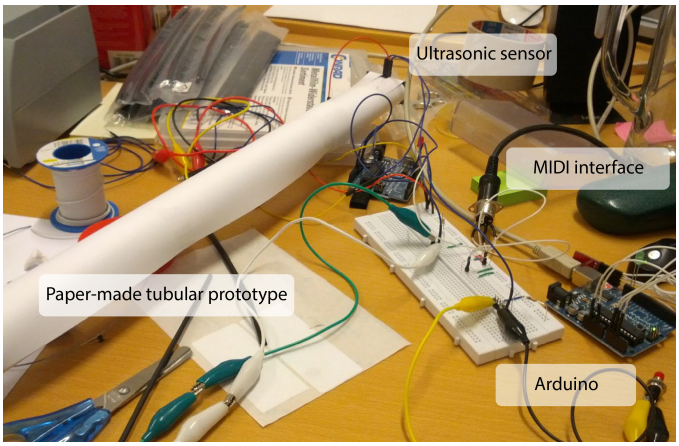


Figure 4: The initial prototype development setting.

During some prototype test sessions, we figured out significant usability problems with the paper tube. In addition, most of the electronic parts were considered to be attached at the hand-held interface. However, this would have been a weight problem for the paper tube and an aesthetic disturbance due to the anticipated size of the Arduino, the battery pack and other required components. We thought about improving the tube prototype by using stronger material such as aluminium and attaching a small box for the electronics. In the end we decided to leave off the tube entirely which lead the design close to final prototype.

B. Final Prototype

The Trombosonic's hand-held interface is purely electronic without any loose or moveable parts. It is held in one hand, either left or right, with a pinch grip. For data processing it uses an Arduino as described earlier in the design process. An attached RedFly WiFi-Shield [24] sends sensor data as OSC messages wirelessly to a computer running Max/MSP for sound synthesis in our particular case or any other OSC-compliant musical application.

The casing of the interface is cylindrical with rounded ends and made of polystyrene and wood (see Fig. 1 and Fig. 5). This keeps it lightweight but stable and handy. All electronic devices are bolt-on or glued. Additionally, four aluminium rods provide a good grip and they round out the overall appearance. Its total weight including batteries is 294 g (10.37 oz).

For powering the Arduino, a battery pack is included at the bottom of the interface which holds four standard AA batteries. The longest period of time that the Trombosonic was turned on for testing purposes was 130 minutes and no energy problems were observed during this time. An accurate test regarding energy consumption has not been done yet.

Both the compact design and the wireless network communication ensure free and easy movement during usage within the range of the arm and without being wired to the computer. The whole set of sensors and why they are specifically used to enable embodied musical expression, are described in detail in the following.

V. Musical Expression

The Trombosonic's primary intention is to serve as a musical instrument. Hence, it has several features that enable

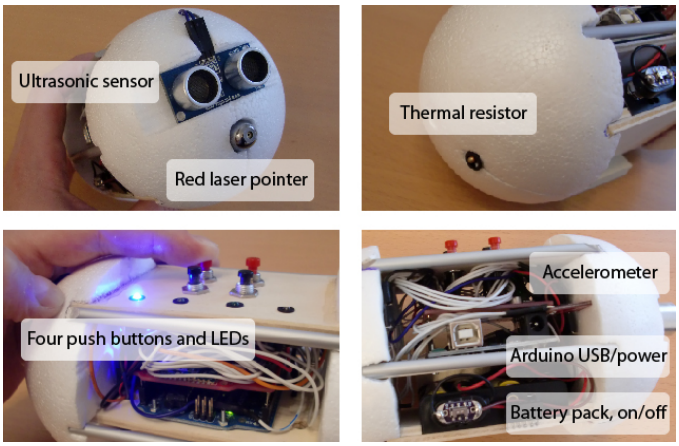


Figure 5: Description of all features of the interface

TABLE II: Overview of features and functionality

Feature	Functionality
Four buttons	Tone on/off, set synthesizer param.
Ultrasonic sensor	Pitch/frequency control
Red laser pointer	Direction of the ultrasonic sensor
Accelerometer	Position/movement of the interface
Thermal resistor	Using the player's breath

expressive sound control, including (1) four push buttons and four LEDs, (2) an ultrasonic sensor, (3) a red laser pointer, (4) a thermal resistor, and (5) an accelerometer. For its use as a musical instrument the sensors and actuators are configured to work for particular musical purposes. All features are shown in Fig. 5 as they are located on the device and an overview with a short description is given in Table II.

A. Physical sound generation

Four push buttons, mounted on the top board, enable the control of the basic functions. They are ordered in a square and operated with the middle finger and the ring finger. This allows a good grip using the other two fingers while pushing the four buttons. For additional visual feedback each button is connected to an LED in a different colour, which flashes when the button is pushed. All combinations of how the buttons can be pushed and the corresponding functions are shown in Table III.

TABLE III: Summary of button functions

Buttons pushed at once	Functionality
1	Tone on/off
2	Set new frequency
3	Switch oscillator wave
4	Switch LFO waveform
1 + 2	Switch filter type
3 + 4	Laser on/off
1 + 2 + 3 + 4	Set thermal resistor value
other button combinations	not used yet

The Trombosonic uses a subtractive sound synthesis. The default frequency of the oscillator is 440 Hz. Button 1 turns on the sound, while button 2 allows the user to save and hold the actual frequency which changes continuously according to play. With this function, the player is able to explore the acoustic frequency spectrum endlessly or at least within the human acoustic range. Buttons 3 and 4 switch between waveforms of the oscillator and a Low Frequency Oscillator (LFO). Pushing buttons 1 and 2 or 3 and 4 together switches between filter types and turns on the laser. Pushing all buttons at once, resets the reference value of the thermal resistor. All functions are described later in detail.

B. Embodied expression

The design of Trombosonic enables a range of embodied expressions in play. The **ultrasonic sensor** [25] at the front enables the typical pitch control of the generated tone as known from the slide trombone. Unlike the traditional instrument the Trombosonic has no slider or handle. Instead, the **red laser pointer** (a disassembled off-the-shelf model for presentations) indicates the direction of the ultrasonic sensor for a better orientation of the pitch-steering hand as shown in Fig. 3. While moving it back and forth a red dot is projected on the palm. This realisation allows the player to play the instrument with two hands, comparable to a slide trombone which also makes it familiar to spectators in its embodied movements.

Because the ultrasonic waves can bounce off any object, the second hand is not mandatory, thus the Trombosonic can also be played with just one hand and interact with other objects. These objects may be items within the performer's environment, or the body itself. Whatever interface is pointed on, the distance is transformed into sound. Even spectators who are moving or waving hands can allow interactive sonification of both performer and audience. The laser pointer can also be turned off and on at any time during a performance to avoid dazzling the spectators.

Another embodied sound control is realised with an **accelerometer** [26] that measures the interface's movement in three dimensions. The actual synthesizer implementation uses two of them. The device can be turned around the longest axis (the one the red laser points to) and up- and downwards to control frequencies of the LFO and the filter.

Given that the trombone, the source of our inspiration, is a wind instrument, we also included a mouth piece in our interface. Unlike the slide trombone, it is for additional expression only and not the origin of the tone. For reasons of simplicity we did not use a complex breath analyser [5], [6] but a simple **thermal resistor** [27] to recognise the player's breath. During the design process we used this value to intensify different parameters of the synthesizer, such as the bandwidth of the frequency filter. However, with the actual prototype, the breath control gives the volume a boost as this seems to be comparable with a traditional wind instrument.

C. Sound synthesis

For our applications, the Trombosonic uses Max/MSP as control and sound generation unit. The full patch in presentation mode is shown in Fig. 6. Usually, during playing the instrument, the whole patch is controlled remotely with the wireless interface and receives nine different sensor-values (see Table II). These values are received and visualised in the subpatch "Trombosonic interface" as shown in the green area of Fig. 6 on the left.

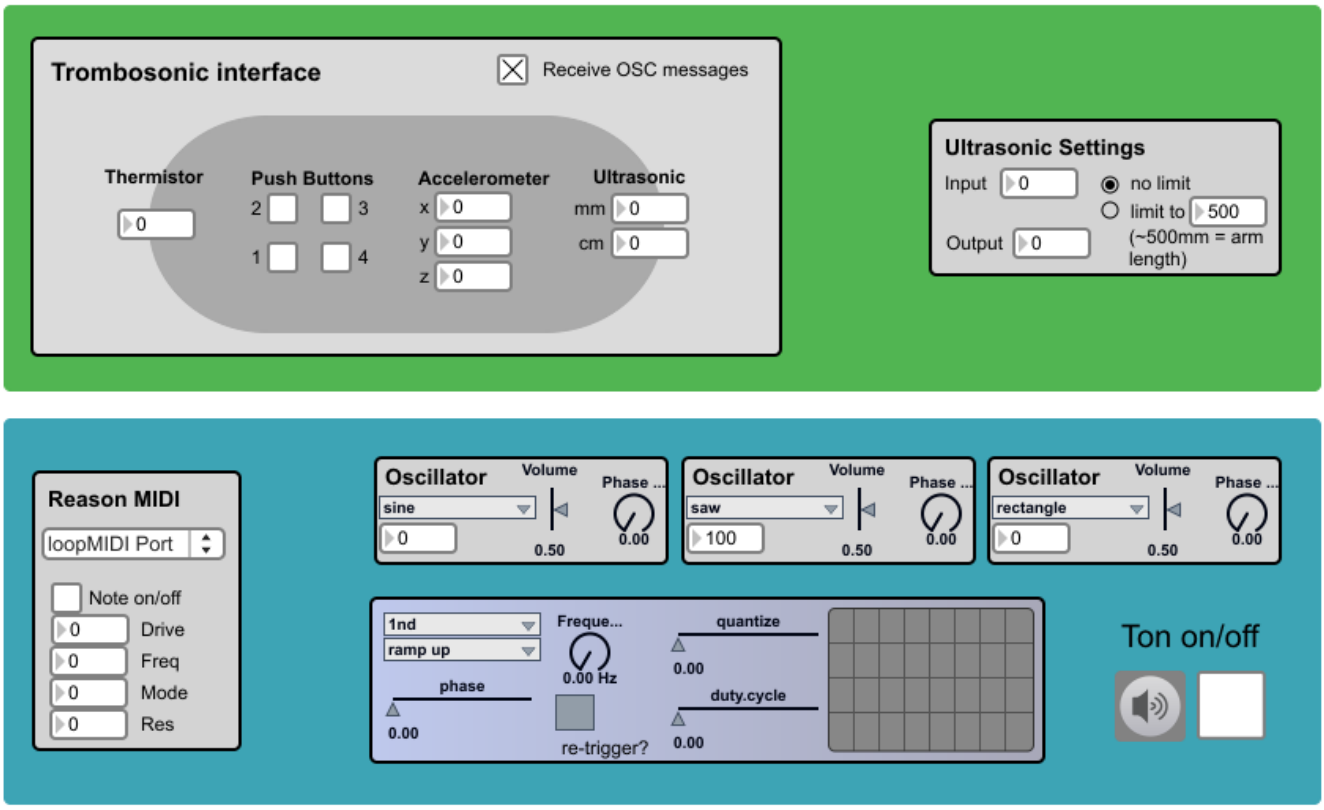


Figure 6: The interface for the Trombosonic sound synthesis as Max/MSP patches in presentation mode

Another sub-patch, "Ultrasonic Settings", allows the player to switch between two modes: (1) no limit or (2) limit to certain distance. If no limit is set, the ultrasonic sensor is capable of measuring a distance of about 4 m. This mode is necessary, if the Trombosonic is played without the second hand. Otherwise it can be limited to the arm length of about 50 cm.

To generate a sound, the player has two opportunities, which are located in the blue area: either to use the built in synthesizer or use the external synthesizer. The internal synthesizer uses subtractive sound synthesis and its sub-patches are located in the lower right of Fig. 6. The player can choose among simple waveforms of three oscillators which are attenuated by ADSR (Attack Decay Sustain Release) envelopes, an LFO and filter effects. Certain parameters can be controlled in real-time with the Trombosonic's hand-held interface.

For this prototype we focused mainly on the interface and minimised the synthesizer's features. Hence, the sound reminds one a little of old synthesizers. Furthermore, there is no special musical training or knowledge needed to play the Trombosonic and to explore its features.

The second option for sound generation is an external synthesizer. In our case, the patch is prepared to be used with Propellerhead's Reason [28] through internal MIDI which is realized by the sub-patch in the lower left of Fig. 6. In our particular setting, the note on/off and four parameters can be controlled remotely with the Trombosonic's hand-held interface.

After the detailed description of the Trombosonic's functionality and its possibilities for musical expression, we proceed with the presentation of the exploratory evaluation.

VI. Exploratory Evaluation

We did an informal evaluation of the Trombosonic as a musical instrument and explored the potential applicability of our prototype in different fields. For this purpose we asked experts at a competition for new musical instruments and a researcher in game and interaction design.

Additionally we build on knowledge from existing literature about the value of extreme users [7]. The positive impact of music and the suitability of musical instruments in various non-music domains have already been shown. For instance they can be used as playful, toy-like devices for non-specialists (e.g., [4], [20]) or for therapeutic prevention and rehabilitation (e.g., [21], [22]). This inspired us not only to use expert knowledge to evaluate the Trombosonic but also to give it away to people with different abilities and ages such as a 92-year-old woman and a 13-year-old boy. We considered them as untypical users for new musical interfaces and expected them to help to explore the Trombosonic's potential beyond performances.

Everyone participating in the exploratory evaluation was not involved in the project before and saw the device for the first time. After a short introduction they were allowed to play the interface freely. Afterwards they were asked to tell us about their experience. In addition, we took photographs of their explorations and took notes of their comments. All subjects started to play the Trombosonic with its general sonic and gestural features we described in Section V. With each of them we spent about 30-45 minutes for exploration and talked with

them about their experience. For some we slightly changed single features tailored to their anticipated interests and needs, as will be described below.

A. New interface for musical expression - expert evaluation

To hand out the Trombosonic to a musician is the most obvious test for a musical instrument. We did so in early spring 2013 for a performance in Vienna where the Trombosonic was used as special instrument for a certain part of a show. The artist used it as a solo instrument during one song.

However, to "fit" better with the other unusual users and to get the most interesting and diverse results, we did something different. We applied for the annual Margaret Guthman Musical Instrument Competition [29], which is considered one of the largest competitions for new musical instruments. The Trombosonic was chosen out of more than 70 submissions to take part in a performance as one of 17 semifinalists which means an acceptance rate of lower than 25%. The successful submission to this highly competitive and renowned competition proves that the Trombosonic is already well-regarded as a new musical instrument. The actual performance took place in Atlanta, USA, in April 2013. We took advantage of this event to get the official feedback of the expert jury as well as the opinion of other participants and audience members when presenting it as a new musical instrument.

The performance at the competition was successful and two pieces were presented: One original electro-acoustic composition and one rather mainstream oriented piece accompanied by pre-recorded playback. People in the audience as well as the jury enjoyed the presentation of the many different features and how the Trombosonic was played in a trombone-like manner during the first piece. The second piece was called "Trombopolka" and was intended to be a tribute to the original instrument, the slide trombone. The Polka is a popular genre of folk music. Some audience members explicitly stated after the performance how they liked the combination of traditional music and the new musical instrument.

The experts mainly criticized deficiencies in the sound synthesis and some spectators missed the acoustic traceability of the breath sensor. One suggestion from another musician was to integrate a microphone for additional sound creation using the human voice. Two other performers pointed out the compact and wireless design, which makes it easy for embodied performances as they anticipated.

In summary, the performance at the musical instrument competition confirms the potential of the Trombosonic as a new interface for musical expression and various comments from new musical instrument experts suggest the direction of future revisions and improvements.

B. Physical training for older adults

We then gave the Trombosonic to a 92-year-old woman who is a relative of one of our project members. She was willing to help us for the evaluation during a visit at her own house. She has full mental abilities apart from some forgetfulness from time to time, as she confessed herself. She is still able to walk without a cane in her home. She told us, she uses a walking stick only outside as a precaution and especially during the winter season. However, according to her own description her movement abilities are getting worse and her visibility is already in a bad condition. Asked for her musical knowledge she said, she had learned to play the piano a long time ago and loved to play music and to sing. Now she is unable to play any more since she cannot see the keys and the score.

Figure 7: A 92-year-old woman playing the Trombosonic: First impression (left), standing to operate it "in another way" (right)

We did not present the Trombosonic as a music instrument to her. According to what literature suggests in relation to physical activity and elderly people [30] we rather said it was an acoustic training device. Addressing her own musical experience, we changed the original electronic sound with a piano synthesizer to make it sound more familiar to her. After an explanation of the buttons and some possibilities to make sound, she started to handle it by herself.

Conversation with her and our own observation have shown that the originally intended way to play the Trombosonic with two hands was not very convenient for her. What was notable though, was her behaviour changing her hands holding the device alternately in both hands and finally she even stood up to operate it "in another way" as she noted (see Fig. 7, right). She said she tried to find a good way to hold it and at the same time preventing her arms from getting tired when moving the device by changing hands. Unlike all other participants of the exploratory evaluation, she was the only one considering tiring issues during playing the instrument. This might be important when using the Trombosonic for older adults or rehabilitation.

It appeared to us that she mainly concentrated on the device itself instead of producing particular sounds. However, at the end of our session she summarized her experience: "I really enjoyed making it sound like a piano doing moves I am usually not used to do. Though I do not know how it works and why it sounded like a piano" (Translated from German).

Overall, we identified a certain interest in the Trombosonic and her different ways to handle it. Following Rolland et al. who illustrate that "regular physical activity is a key component of successful aging" [30] and Bruhn and Schröter who discuss the positive impact of making music in old age, we propose the Trombosonic as a potential device for elderly people. It might be a good way to combine physical and musical activity.

C. Playful interface for children

When talking about musical play and young children, Tarnowski explains "functional musical play might include exploring vocal, instrumental, and environmental sounds as

Figure 8: A boy aged thirteen explores the Trombosonic's features

Figure 9: Playing the Trombosonic as a one-handed instrument to acoustically explore a shelf

well as the way in which these sounds are made" [31]. This motivated us to give the Trombosonic to a young boy aged thirteen (Fig. 8), who was visiting our lab for a trial internship. He has no instrumental training but considers himself a very interested listener to music, which is also indicated by the big headphones he wears around his neck all the time. Additionally, he started to make music with his computer a little while ago, experimenting with a software-synthesizer.

Similar to the older adult, we explained the basic functionality of the Trombosonic to him and how to handle it. When he started playing we observed, most different to all other evaluation participants, that he really seemed to focus on the music. We also noticed that he played the Trombosonic mostly in its originally intended way using two hands. However, once he started to roll the interface with one hand on the table to create a smooth wave-like sound using the accelerometer. He was the only one who used the movement features of the interface in this physical way together with other objects such as the table.

In all, the young boy carefully analysed the different features and ways to play the Trombosonic throughout his whole session. Following his own words "it was a lot of fun" and he would like to control his own sounds with the interface. We propose the Trombosonic as a suitable instrument for letting young people playfully explore music without being able to play a traditional instrument.

D. Sonification and people with disabilities

Finally, we asked a researcher with expertise in interaction design within our lab to tell us about his experience with the Trombosonic. After an initial explanation of the basic functionality we let him explore the device. It was significant that he started to use it as a one-handed device despite our initial advice to play it in a trombone-like manner. Following his own "intuition" (as he defined it by himself) he started to walk around the room using the Trombosonic as a sonification device. He started to explore the environment acoustically while pointing the device onto different walls and surfaces (Fig. 9).

Furthermore, he turned the device around pointing the ultrasonic sensor towards his own body. This way to play the Trombosonic is illustrated in Fig. 10. Moving it back and forth he started explaining: "Look, now it is a one-handed instrument. I can play a trombone without my second hand" (Translated from German). During his test he complained about the lack of clear feedback when using the buttons to control the synthesizer. Since he was not familiar with the synthesizer's options this was really a problem when trying to intentionally switch between wave forms and filters as he said.

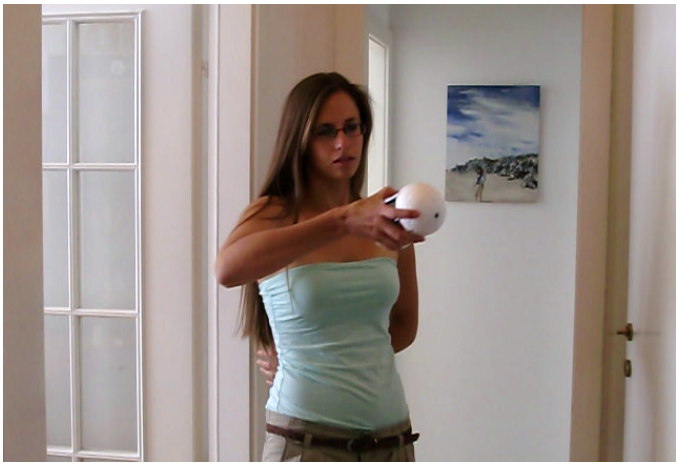


Figure 10: Playing the Trombosonic as a one-handed instrument towards the own body

The trial with the interaction designer suggests some usability improvements for a more intuitive handling. Furthermore, it might be worth considering the Trombosonic as a one-handed musical instrument keeping in mind that "thousands of people with disabilities in the UK, and millions across the world, are excluded from music making" [32]. The Trombosonic could be such an instrument to enable those people and people with restricted mobility in general to gain a trombone-like musical experience. It could also have potential for people with visual impairment as a way to playfully explore their physical environment.

E. Discussion

The exploratory evaluation was not meant to be comprehensive but to complement the main contribution of this paper, the presentation of a new musical instrument. It gave us a differentiated impression of how people play the Trombosonic from the perspectives of both experts (in new musical instruments and in interaction design) and extreme users (very old, young). Their considered feedback, as well as their unanticipated uses, pointed to potential applicability that might be worth considering and gave some initial directions for future development.

Overall, people tried various different ways to handle the interface, such as using one-hand only or both hands and while standing, sitting or moving around. The actual usage to produce sound was ranging from playing music following scores in a traditional way to acoustically explore the environment.

The approach of using an exploratory evaluation when testing new musical instruments turned out to be qualified. It was inspiring to use expert knowledge as well as to see unexpected behaviour of unusual users. We argue that our assumption to widen the range for non-obvious applications by doing an exploratory evaluation was verified. At least for the initial of a new musical instrument and a new interaction device this opened a set of unpredictable possibilities for improvements and new directions to focus on for future development.

Compared to the initial described approaches which augment traditional trombones or create new interfaces on the basis of the original instrument, our strategy to create a compact device has its advantages as the exploratory evaluation has shown. Despite its different appearance, people considered the Trombosonic to be a trombone-like instrument. At least when it is played as intended which happened during the instrument competition in our particular case. Analogously, people tend to play with the interface in unusual ways and they explore its features as soon as they do not think of it as a trombone-like instrument such as the interaction designer and the older adult.

Thus, designing a new musical instrument under certain constraints and evaluate the prototype in an exploratory manner brought the anticipated insights in unexpected and unpredictable user behaviour. The combination of experts and extreme users helped to go beyond the usual applicability of this musical interface in fields such as healthcare and education.

VII. Evaluation at a Participatory Performance

Guided by the results of the exploratory evaluation, we improved the quality of the sound capabilities and conducted a more focused, music-oriented evaluation to explore the experience of the audience who attends a Trombosonic performance. This study did not fully work out as intended as we will describe later. However, there was still enough data to draw out directions for future in-depth studies.

We used a public live concert of the first author's band "Oliver Linus" at the festival "Wiener Musik-Experimente" (Viennese Music-Experiments) on 6th February 2014 in Vienna, Austria. This was a good opportunity for us from a conceptual point of view in particular, because the main idea behind this event was to interlink mainstream and experimental approaches in live music in various different ways.

Moreover, the band already planned to do a participatory performance for other research purposes at this concert. The intention was to let the audience influence certain sound effects in real-time during the performance of one piece by using a big balloon. This was a central element of the whole show. The results of that particular study are discussed elsewhere [8].

The main motivation for the Trombosonic part of the performance was to look more deeply at the potential of the Trombosonic as a music instrument and to get insight into the audience's experience and opinion about it. To not interfere with the other study and overwhelm the audience, we decided to change two things significantly. First, the Trombosonic performance should be a solo without the band. Second, since the audience participation of the other study was in real-time during the show, the Trombosonic performance should contain asynchronous participation immediately after the concert in the form of active experience collection among the audience.

The whole performance of the band "Oliver Linus" took about 30 minutes. From a conceptual point of view and the first author's performance perspective, the show was in three parts: (1) Two solo pieces with the Trombosonic, (2) two regular songs with the band, and (3) two songs including the audience for real-time participation. A visual impression of the setting is documented from three perspectives in Fig. 11. The whole Trombosonic performance including two pieces is available online as a video [33].

The reason for choosing two songs for the Trombosonic performance was on purpose. The first one was the electro-acoustic piece "From Peak to Sine" and the second one was called "The night the stars fell asleep" which can be considered as popular music. Both songs were instrumental and the second one had an accompanying playback. This setting should give us the chance to gather the experience of the audience with the Trombosonic using the example of two very different songs.

The original idea was to let the audience participate asynchronously and decide right after the show which of the two Trombosonic songs should be played again as an encore. By doing so, we wanted to get the preference of the audience regarding to the two very different Trombosonic songs. We thought about using applause and cheering as an instant measurement similar to other studies (e.g., [19]). Finally, this did not work out at all for organisational reasons mainly. The whole festival was far behind the time schedule and so there was simply not enough time for another song and the decision making process of asking the audience again to applause and cheer for the two songs. Unfortunately, we were also not prepared to do the measurement with technical means right after the songs' performance.

However, there was still another chance to gather data about the Trombosonic performance. This data was collected with the help of a short questionnaire for evaluation right after the show during the preparation of the stage for the next band. The questionnaire itself and its results are presented in the following.

A. The After-Concert Questionnaire

After the concert, 32 out of approximately 80 spectators were randomly selected and asked 10 questions about their experience during the show of the band "Oliver Linus". The main purpose of this questionnaire was the real-time participation of the audience for another study [8] as already mentioned. Hence, most of the questions were aimed towards the experience during the interaction with the balloon. Nevertheless,

Figure 11: Three perspectives of the performance at the music festival "Wiener Musik-Experimente": The performer playing the Trombosonic "against the wall" (top), the audience watching the performance (middle), and the view from the back towards the stage (bottom)

three questions were targeting the whole show[1]:

1) Describe your impression with one or two words?
2) What did you like best?
3) What did you dislike?

We consider these general questions as partly relevant for the Trombosonic as some participants mentioned it explicitly in their answers. For the analysis we coded all contents of the relevant three questions thematically.

All answers of the first question, which are basically only single words, were categorized as either *positive*, *neutral* or *negative*. The number of words per participant varied from 0 to 3. This resulted in a total number of 39 words that describe e.g., a feeling, an opinion, or an experience. The choice whether a word was rated as positive, neutral or negative might be ambiguous. No word was coded twice and all words which were considered as not clearly positive or negative were counted as neutral. The summary of this analysis is presented in Table IV.

[1]All contents related to the questionnaire were translated from German to English

TABLE IV: Summary of the thematic analysis of question 1

Code	Number of words	Examples
Positive	18	atmospheric, rousing, funny, super, inspiring
Neutral	15	interesting, different, technical, electric
Negative	6	confusing, nervous, unpleasant, tedious

In the second and third relevant question, people had to say what they like best and dislike. Here, the Trombosonic in particular was mentioned twice and once the overall performance experience was commented. Since the only electronic piece was one played with the Trombosonic, we considered these statements to be relevant here as well. The selected answers according to the questions are listed in Table V in the original language (German) and the English translation.

TABLE V: Summary of the answers to questions 2 and 3 that are relevant for the Trombosonic

English translation	German original
Q: What did you like best?	
Trombosonic was cool, singing	Trombosonic war cool, Gesang
Trombosonic, Music (not electronic)	Trombosonic, Musik (nicht elektronisch)
Was funny, interesting overall exp.	War lustig, interessantes Gesamterlebnis
Q: What did you dislike?	
first Trombosonic song	erste Trombosonic Lied
electronic music	elekronische Musik

B. Discussion

The evaluation of the Trombosonic at a participatory live performance by doing a short survey right after the concert, has raised some indications that could be considered for further investigation in future studies.

When asked for single words to describe the experience of the performance, the majority of all statements (33 out of 39) were positive or neutral. Some were more general such as *funny* or *super* but others gave descriptions such as *atmospheric*, *rousing* or even *inspiring*. Spectators who had negative experiences found the performance *confusing*, *nervous* or *tedious*. Even one spectator rated the show as *unpleasant*.

However, none of these experiences really addresses aesthetic or music issues. Apart from *atmospheric* and maybe *electric*, all statements indicate a certain experience associated with a feeling.

This was different when the spectators were asked about what they liked or disliked. Here in four of all five statements, where the Trombosonic was mentioned, participants talked about music-relevant issues. Notable in this case was that two people talked about the *electronic music* in a negative context. One even said "Trombosonic, Music (*not electronic*)" when asked about the positive experience. This indicates, that the instrument itself is perceived separately from the music.

This can be interpreted differently. If anticipated that spectators "accept" or even "like" a new digital music instrument

in principle, there is still the question of "how it is played" and "what is played" which makes the overall experience. In our particular study, we performed two totally different kinds of music. One pure electro-acoustic piece with a simple synthesizer and one popular-music oriented piece with a full arrangement. At this point we can only anticipate that the people preferred the second piece for aesthetic reasons while they found the whole instrument *interesting*, *funny* and *inspiring*.

After presenting the two evaluation studies, one explorative with extreme users and one at a concert with a participatory performance, we proceed with the conclusion of these insights.

VIII. Conclusion

The Trombosonic is a new instrument for musical expression that derives from the slide trombone. However, it does not imitate the slide trombone either visually or acoustically, rather the principles of this wind instrument serve as a design inspiration for the interactive gestures.

Push buttons, an ultrasonic sensor and a red laser allow an embodied playing of the instrument similar to the slide trombone changing the pitch with one hand moving back and forth. Compared to a traditional slide trombone, the whole instrument's size is much smaller and the slider is completely missing. Furthermore, an accelerometer and a thermal resistor enable an additional embodied expression. Moving the whole interface enhances the musical possibilities compared to the traditional instrument, while the use of the player's breath retains a typical feature of wind instruments.

Along with presenting the Trombosonic as a new interface for musical expression we did an exploratory evaluation looking for its potential as a musical instrument as well as in other fields. Hence, we successfully submitted a performance proposal to an international competition for new musical instruments and gave the instrument to a 92-year-old woman, a 13-year-old boy and a researcher in game and interaction design. This let us identify different issues and unexpected aspects to keep in mind for future improvements. All cases also indicate the Trombosonic's suitability for various musical purposes as well as non-music applications.

In addition to the exploratory evaluation we conducted a typical music-related evaluation for new instruments at a live performance. This was conceptualized as a participatory performance including the audience. After the concert randomly selected spectators were asked with the help of a short questionnaire about their experience. Their statements are not enough evidence to draw definite conclusions but they indicate a certain connection in relation to the perception and experience of a new digital music instrument. We assume a dependency between the overall concept of a new instrument, the way it is played and for what kind of music it is used. Furthermore, in a participatory performance, which has certainly new elements for an audience per se, as it was in our study, it is even more important that the balance between new concepts, new technology and new music is considered for the sake of the audience's overall experience.

IX. Future directions

For advanced prototypes of the Trombosonic we plan to integrate a microphone for additional sound creation using the human voice. Furthermore, the synthesizer needs some revision regarding the sound and better mapping of sensor values to single parameters, along with a more intuitive button control. Beyond technical improvement addressing mainly musical features, the evaluation suggests to adapt and use the interface in other domains. It could be used as training device for elderly people addressing physical and musical health-relevant activity or it could let children intuitively explore sound generation without being trained to play a traditional music instrument.

The field studies, and in particular the evaluation at the participatory performance, have shown the potential of such settings. Future studies might include a more substantial methodology and a focused approach when asking spectators about their experience with a new music instrument. The concept of using very different sorts of music with the same new instrument and the same audience seems promising and could be adapted for an in-depth study about new digital music instruments in relation to experience, opinion, and aesthetics.

References

[1] O. Hödl and G. Fitzpatrick, "Trombosonic: Designing and Exploring a New Interface for Musical Expression in Music and Non-Music Domains," in The Seventh International Conference on Advances in Computer-Human Interactions, 2014, pp. 54–59.

[2] C. Dobrian and D. Koppelman, "The 'E' in NIME: Musical Expression with New Computer Interfaces," in Proceedings of New Interfaces for Musical Expression, 2006, pp. 277–282.

[3] C. Pacchetti and F. Mancini, "Active music therapy in Parkinson's disease: an integrative method for motor and emotional rehabilitation," vol. 393, 2000, pp. 386–393.

[4] D. Robson, "Play!: Sound toys for non-musicians," Computer Music Journal, vol. 26, no. 3, 2002, pp. 50–61.

[5] N. Farwell, "Adapting the trombone: a suite of electro-acoustic interventions for the piece Rouse," in Proceedings of New Interfaces for Musical Expression, 2006, pp. 358–363.

[6] T. Henriques, "Double Slide Controller," in Proceedings of New Interfaces for Musical Expression, 2009, pp. 260–261.

[7] L. Holmquist, "User-driven innovation in the future applications lab," Extended abstracts of the 2004 conference on Human factors and computing systems - CHI '04, 2004, pp. 1091–1092.

[8] O. Hödl, G. Fitzpatrick, and S. Holland, "Experimence: Considerations for Composing a Rock Song for Interactive Audience Participation," in Proceedings of ICMC, 2014.

[9] N. Collins, "Low Brass: The Evolution of Trombone-Propelled Electronics," Leonardo Music Journal, vol. 1, no. 1, 1991, pp. 41–44.

[10] S. Lemouton, M. Stroppa, and B. Sluchin, "Using the augmented trombone in I will not kiss your f.ing flag," in Proceedings of New Interfaces for Musical Expression, 2006, pp. 304–307.

[11] M. A. Bromwich, "The Metabone: An interactive sensory control mechanism for virtuoso trombone," in Proceedings of International Computer Music Conference, 1997, pp. 2–4.

[12] M. Su, W. Lee, and S. Chen, "Electronic trombone: an interactive tool to promote musical learning and performance creativity," in Proceedings of the 10th International Conference on Virtual Reality Continuum and Its Applications in Industry, 2011, pp. 585–588.

[13] C. Kiefer, N. Collins, and G. Fitzpatrick, "HCI methodology for evaluating musical controllers: A case study," in Proc. of NIME, 2008.

[14] D. Stowell, A. Robertson, N. Bryan-Kinns, and M. Plumbley, "Evaluation of live humancomputer music-making: Quantitative and qualitative approaches," International Journal of Human-Computer Studies, vol. 67, no. 11, Nov. 2009, pp. 960–975.

[15] E. W. Pedersen and K. Hornbæ k, "mixiTUI: A Tangible Sequencer for Electronic Live Performances," in Proceedings of the 3rd International Conference on Tangible and Embedded Interaction - TEI '09. ACM Press, 2009.

[16] G. Kaiser, G. Ekblad, and L. Broling, "Audience Participation in a Dance-Club Context: Design of a System For Collaborative Creation of Visuals," in In Proceedings of Nordes, 2007.

[17] S. Thompson, "Audience responses to a live orchestral concert," Musicae Scientiae, vol. 10, no. 2, Sep. 2006, pp. 215–244.

[18] R. Knapp, J. Jaimovich, and N. Coghlan, "Measurement of motion and emotion during musical performance," in Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on. IEEE, 2009, pp. 1–5.

[19] L. Barkhuus, "Engaging the crowd: studies of audience-performer interaction," CHI '08 Extended Abstracts on Human Factors in Computing Systems, 2008, pp. 2925–2930.

[20] S. Jordà, "Digital Instruments and Players: Part I Efficiency and Apprenticeship," in Proceedings of New Interfaces for Musical Expression, 2004, pp. 59–63.

[21] C. Plahl, "Musiktherapie - Praxisfelder und Vorgehensweisen," in Musikpsychologie. Enzyklopädie, Rowohlts, 2009, pp. 630–652.

[22] M. E. Clark, A. W. Lipe, and M. Bilbrey, "Use of music to decrease aggressive behaviors in people with dementia." Journal of gerontological nursing, vol. 24, no. 7, Jul. 1998, pp. 10–17.

[23] Arduino. [Online]. Available: http://www.arduino.cc (last access 2. Sept. 2014)

[24] RedFly WiFi-Shield. [Online]. Available: http://www.watterott.com/de/Arduino-RedFly-Shield (last access 2. Sept. 2014)

[25] Ultrasonic sensor. [Online]. Available: http://www.watterott.com/en/Parallax-PING-Ultrasonic-Sensor-28015 (last access 2. Sept. 2014)

[26] Accelerometer ADXL335. [Online]. Available: http://www.watterott.com/en/Breakout-Board-ADXL335 (last access 2. Sept. 2014)

[27] Thermal resistor. [Online]. Available: https://www.sparkfun.com/products/250 (last access 2. Sept. 2014)

[28] Reason. [Online]. Available: https://www.propellerheads.se/products/reason/ (last access 2. Sept. 2014)

[29] Margaret Guthman Musical Instrument Competition. [Online]. Available: http://www.gtcmt.gatech.edu/guthman2013 (last access 2. Sept. 2014)

[30] Y. Rolland, G. Abellan van Kan, and B. Vellas, "Physical activity and Alzheimer's disease: from prevention to therapeutic perspectives." Journal of the American Medical Directors Association, vol. 9, no. 6, Jul. 2008, pp. 390–405.

[31] S. M. Tarnowski, "Musical Play and Young Children," Music Educators Journal, vol. 86, no. 1, 1999, pp. 26–29.

[32] The One-Handed Musical Instrument Trust. [Online]. Available: http://www.ohmi.org.uk (last access 2. Sept. 2014)

[33] Trombosonic Performance at the festival Wiener Musik-Experimente. [Online]. Available: https://www.youtube.com/watch?v=wm48g4xL8QQ (last access 2. Sept. 2014)

Extended Trace-based Task Tree Generation

Patrick Harms, Steffen Herbold, and Jens Grabowski
Institute of Computer Science
University of Göttingen
Göttingen, Germany
E-mail: {harms,herbold,grabowski}@cs.uni-goettingen.de

***Abstract*—Task trees are a well-known way for the manual modeling of user interactions. They provide an ideal basis for software analysis including usability evaluations if they are generated based on usage traces. In this paper, we present an approach for the automated generation of task trees based on traces of user interactions. For this, we utilize usage monitors to record all events caused by users. These events are written into log files from which we generate task trees. The generation mechanism covers the detection of iterations, of common usage sequences, and the merging of similar variants of semantically equal sequence. We validate our method in two case studies and show that it is able to generate task trees representing actual user behavior.**

Keywords-task tree generation, usage-based, traces, task tree merging.

I. Introduction

Task trees are an established method to model user interactions with websites. They can be created manually at design time or automatically, e.g., based on recorded user actions [1] or based on existing Hyper-Text Markup Language (HTML) source code [2]. When created manually at design time [3], the structure of task trees reflects the user interactions as intended by the interaction designer [4]. Task trees can also be used for comparing expected and effective user behavior as a basis for a semi-automatic usability evaluation [5]. When they are not generated based on recorded user actions, task trees do not describe effective user behavior but either expected or possible user behavior.

In this paper, we present an approach for automatically generating task trees based on recordings of user interactions. This approach does not require a manual marking of task executions in the recorded data before the task generation making it easily applicable to larger data sets and differentiating it from other approaches. The generated task trees represent the effective behavior of users. They can, therefore, be used for a detailed analysis of the usage of a software what is the major goal we intend to achieve based on our approach. For example, in other work we utilize the generated task trees for an automatic usability evaluation of a website [6]. The results of a usage analysis are used for optimizing software with respect to the user's needs. Throughout the remainder of this paper, we use the analysis of websites as a running example. However, our approach is designed for event-driven software in general including desktop applications.

The approach in this paper is an extension of our work described in [1]. The extension covers mainly the detection and merging of similar generated sequences. We provide details about the challenges introduced by the merging process and extended the case studies section to also evaluate the merging results. Furthermore, we added sections describing in more details why and how the recorded user actions need to be post-processed and how we detect and handle common elements on different pages of a website, e.g., menu structures, to improve the detection of tasks. Finally, we extended the related work section to compare our approach and the resulting task trees in more details with other approaches.

The remainder of this paper is structured as follows. First, we introduce our approach and the respective terminology in Section II. Then, we describe an implementation in Section III. In Section IV, we present two case studies in which we tested the feasibility of our approach and discuss the case study results in Section V. Finally, we refer to related work in Section VI and conclude with an outlook on future work in Section VII.

II. Trace-based Task Tree Generation

The goal of our approach is to generate task trees based on recorded user actions. In this section, we introduce the details of the approach that consists of four major steps: user interaction tracing, data preparation, detection of sequences and iterations, and merging of similar sequences. These major steps are shown in Figure 1. We commence with the definition of terms that we use in this paper. In the subsequent sections, we describe the details for the four major steps.

A. Terminology

Users utilize a website by performing elementary *action*s. An action is, e.g., clicking with the mouse on a button, typing some text into a text field, or scrolling a page. Actions cause *event*s to occur on a website, also known as Document Object Model (DOM) events. An event is characterized by a *type* that denotes the kind of event and, hence, the type of action that causes the event. Furthermore, an event refers to a *target* indicating the element of a website on which the corresponding action was executed and where the event was observed. For example, clicking with a mouse on a link (action) causes an event of type `onclick` with the link as its target. Typing a text into a text field causes an event of type `onchange`

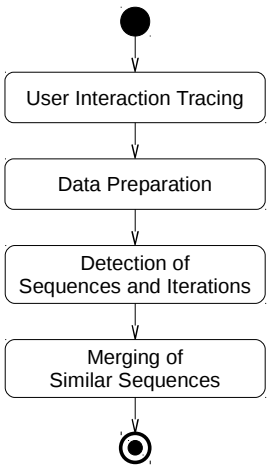


Figure 1. Overall process for generating task trees

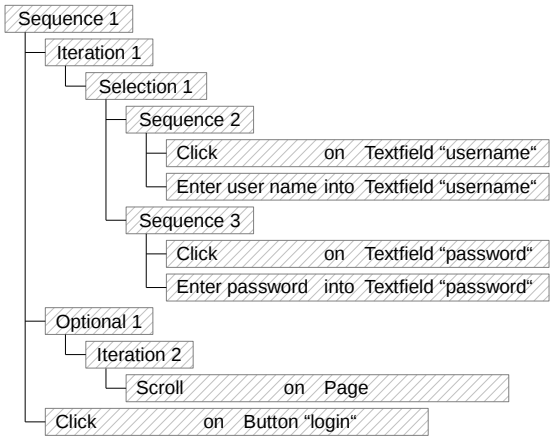


Figure 2. Example for a task tree

with the text field as its target. Events are representations of actions. For each action there is a mapping to an event caused by performing the action.

To execute a specific *task* on a website, users have to perform several actions. For example, for logging in on a website, users must type in a user name and a password into two separate text fields and click on a confirmation button.

Tasks and actions can be combined to form higher level tasks. For example, the task of submitting an entry on a forum website comprises a *subtask* for logging in on the website as well as several actions for writing the forum entry and submitting it. Therefore, tasks and actions form a tree structure called a *task tree*. The leaf nodes of a task tree are the actions users must perform to fulfill the overall task. The overall task itself is the root node of the task tree. The intermediate nodes in the task tree define a structure of subtasks for the overall task.

A task defines a *temporal relationship* for its children, which specifies the order in which the children (subtasks and actions) must be executed to fulfill the task. Different task modeling approaches use different temporal relationships [7]. In our work, we consider the temporal relationships *sequence*, *iteration*, *selection*, and *optional*. If a task is a sequence, its children are executed in a specified order. If a task is an iteration, it has only one direct child, which can be executed one or more times. If a task is a selection, only one of its children can be performed. If a task is an optional, it has only one child whose execution can be skipped. A leaf node in a task tree has no children and does, therefore, not define a temporal relationship.

An example for a task tree is shown in Figure 2. It represents the actions to be taken to perform a login on a website. The actions are the leaf nodes. The temporal relationships of their parent nodes define the order in which the actions have to be performed. The task starts with an iteration of a selection. The possible variants are entering a user name or a password in the respective fields. Users may enter and change their user name and password several times. Optionally, users scroll before completing the overall task by clicking the login button.

The execution of a task is called a *task instance*. A task instance is a tree structure similar to that of the corresponding task. It reflects in detail how a task and its subtasks were executed. Each node in a task instance refers to the task of which it is an instance. The concrete structure of a task instance depends on the temporal relationships of the corresponding task. For example, an iteration has only one child, but an instance of an iteration (i.e., an *iteration instance*) has as many children as the single child of the iteration was executed. In contrast, a selection has several children but a *selection instance* has only one child that is an instance of the executed child, i.e., the chosen variant, of the selection.

An example of a task instance for the task tree in Figure 2 is shown in Figure 3. For executing the task *Sequence 1*, first the *Iteration 1* is executed two times indicated by its two children. Both children are instances of the child of *Iteration 1*, i.e., *Selection 1*. In the first instance of *Selection 1*, the user selected *Sequence 2* to be executed, in the second instance, *Sequence 3* was performed. To finalize the instance of *Sequence 1*, the user did not scroll before clicking the login button, which is indicated by an instance of *Optional 1* having no child instance.

A simplified representation of a task instance is a *flattened task instance*. A flattened task instance is an ordered list of the actions that were executed. This is identical to listing the leaf nodes of a task instance.

B. User Interaction Tracing

The first step in our approach is tracing user actions on a website. This is done by recording the events caused by the actions of users. We achieve this by integrating a monitoring module into the website. This module is invisible to the users and has minimal effect on the implementation, performance, and stability of the website [8]. The module can also be configured to ensure privacy by dropping user data from the recorded events. The resulting sequence of events is encrypted, sent to a server, and stored in a log file. A recorded sequence of events is called a trace.

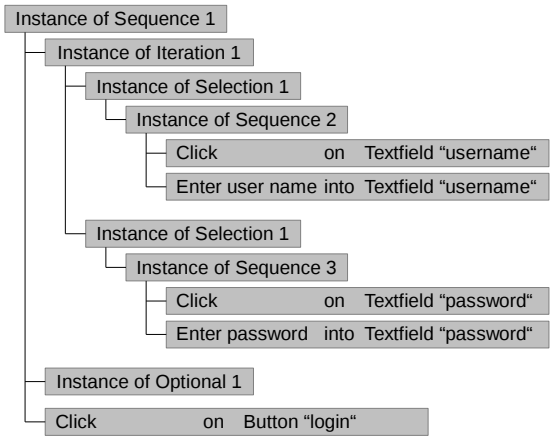


Figure 3. Example for an instance of the task tree in Figure 2

1.	Left mouse button click	on	Textfield with id username
2.	Text input „usr“	on	Textfield with id username
3.	Left mouse button click	on	Textfield with id username
4.	Text input „user“	on	Textfield with id username
5.	Left mouse button click	on	Textfield with id password
6.	Text input „“	on	Textfield with id password
7.	Left mouse button click	on	Button with name „login“

Figure 4. Example for a trace

A simplified example of a trace is shown in Figure 4. It lists the events recorded for a login of a single user on a website. The login comprises the entering of the user name and the password in the respective text fields, as well as a confirmation by clicking on the login button. As the user initially entered a wrong user name, it is reentered a second time. The user does not scroll before confirming the login.

C. Data Preparation

The second step in our approach is the preparation of the data gathered from tracing the user interactions. This includes correcting the recorded events and checking the structure of the website for common page elements to identify identical user actions on different pages of a website. We describe this data preparation in the following subsections.

1) Trace Post-Processing: Traces can have structural abnormalities that are unnatural for a user action. These abnormalities can be caused by the high level of detail on which the events are recorded, by the event types, or by the technology used for event recording. A typical example is that a technology provides events indicating a keyboard focus change separately to key strokes. Hence, the target of a key stroke event is influenced by the last preceding keyboard focus change event. In addition, traces may contain several events that together represent a single user action. For example, two subsequent clicks on the same website element in a short time period represent a double click.

Due to these abnormalities, we perform a post-processing of the traces before the generation of task trees. The post-processing can be automated, as the structural abnormalities in the traces always follow a specific pattern. For example, for each key stroke event, we check the last preceding keyboard focus change event and adapt the target accordingly. Afterwards, we drop all focus change events from the traces. Overall, we perform the following post-processing:

- *Detection of double clicks:* Two subsequent click events with the left mouse button on the same website element within a time frame of 500 milliseconds are transformed into a double click event.
- *Correction of the target of key stroke events:* The target of key stroke events is set to the target of the last preceding keyboard focus change event.
- *Correction of tab key navigation:* When navigating from one text field to another, two events are generated: a key stroke event for hitting the tab key and a value change for the first text field. These events are always recorded in reverse order (first tab key stroke, then value change) although logically the value change happened before the tab key stroke. In these cases, the order of both events is switched.

2) Common Page Elements: Nowadays, websites are composed of several pages all having a similar layout. For example, all pages of a website contain the same navigation menu made up of the same links. This means, that specific elements (links, images, buttons, etc.) reappear on all pages of a website. Although invisible for the user, these elements are in fact distinct instances of the same element. We call these *common page elements* as they are common to several pages of a website. On contrary, there are elements on different pages that reside at the same or similar location but are semantically distinct. Examples are form elements of different forms on different pages positioned similarly for design consistency.

When tracing user actions, we respect common page elements. We consider actions on common page elements as identical if they were performed on the same common page element. Otherwise, we consider them as distinct. This is important for the subsequent task tree generation. Without this consideration, generated tasks would be considered distinct although semantically they are not.

Elements on different pages are considered common, if they have the same id. The ids are defined by assigning them to nodes in the HTML Document Object Model (DOM). Usually, this is automatically ensured when using modern content management systems.

If a website does not make use of identical ids for common page elements, we add and harmonize the ids subsequently and automatically. For this we first create a mapping between HTML DOM nodes and the ids they should have. To identify an HTML DOM node, we use the page it belongs to and the complete path of nodes through the DOM of the page pointing to the node. A path through the DOM must be unambiguous. Hence, on every part of such a path, we consider the HTML tag of the node, its id if there is one, and its index with respect to the children of the parent node. An example for such a path is the following:

page1/html[0]/body[0]/table(id=tab_1)/tr[2]/td[3]/img[0]

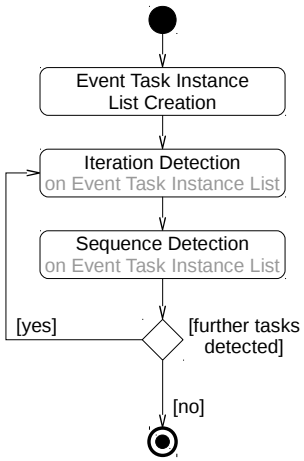


Figure 5. Basic process for detecting sequences and iteration

This path identifies an image embedded in a table on page "page1". The image is located in the table with id "tab_1" in the third row ("tr[2]") and the fourth column ("td[3]"). The image is the first child of the respective table cell.

For each common page element, we create a mapping between the path to the corresponding HTML DOM node and the id it should carry. Afterwards, we update the events in a trace. For each event, we check if the referenced target matches an entry of the mappings. If there is an entry, we update the information about the referenced target to include the id defined in the mapping. To be able to perform a lookup in the mapping, the target references in the events contain all required information including the page the target resides on and its position in the DOM. Eventually, all events having targets with the same id are considered to be recorded on the same common page element.

D. Detection of Sequences and Iterations

The third step of our approach is the detection of sequences and iterations. This step consists of three substeps: event task instance list creation, iteration detection, and sequence detection. The substeps may be executed several times to fully detect all sequences and iterations. The basic execution of these substeps is shown in Figure 5. In the following subsections, we describe these individual substeps and their repeated execution. For this, we first introduce the levels of design, which are important for structuring task trees. We then describe the creation of the initial task trees consisting only of sequences and iterations. Finally, we introduce some characteristics of these initial task trees.

1) Levels of Design: When designing Graphical User Interface (GUI)s, four levels of design are considered: conceptual design, semantic design, syntactical design, and lexical design. They are shown in Figure 6. The conceptual design describes the types of entities that are editable with a software [9], as well as their relationships [10]. For example, in a system for managing addresses, addresses and persons are the entity types. These entity types are related, because a person may be assigned zero or more addresses.

Conceptual Design	Types of entities and their relationships
Semantic Design	Functions to modify entities
Syntactical Design	Steps to take for executing functions on entities
Lexical Design	Physical execution of steps to execute functions on entities

Figure 6. Levels of design

The semantic design specifies functions to edit the entities defined in the conceptual design [9]. For the address management example, this includes adding, editing, and deleting addresses and persons. The syntactical design specifies the steps to execute a function defined in the semantic design [9]. For example, adding a new address is comprised of steps like adding a street name, a city, and a zip code. At the most detailed level, the lexical design specifies means of physically performing steps defined in the syntactical design [9]. In the example, defining a street of an address includes clicking on the respective text field and typing the street name.

In our approach, we map the semantic, syntactical, and lexical levels of design onto task trees. For each function specified in the semantic design, there exists at least one task for executing that function. Hence, there is at least one task tree for each function in the semantic design. The syntactic design is a decomposition of functions into individual steps for function execution. This decomposition corresponds to the definition of subtasks and their temporal relationships within task trees. The actions on the lexical level of design are represented through the leaf nodes of task trees. As we record the events mapped to the respective actions, we refer to the leaf nodes as *event tasks*. Event tasks are considered tasks with the constraint of not having children and not defining a temporal relationship.

2) Event Task Instance List Creation: Using the basic mapping of task trees to the levels of design, we create task trees starting from the leaf nodes, i.e., from the event tasks. For each event in a trace, we generate an instance of an event task. All event task instances are stored in an ordered list. The order in the list is given by the order in which the respective events were recorded. An example is shown in Figure 7a where each grey rectangle denotes an event task instance and the arrows denote their order. The letters of the event task instances denote the respective event task. If this letter is the same for several event task instances, it indicates that the same event task, i.e., the same action, was executed at distinct times.

3) Iteration Detection: The ordered list of event task instances may contain subsequent instances of the same event task. For example, the user might have clicked several times on the same button. Such tasks are represented in task trees as iterations. Therefore, we scan the list of event task instances for iterations of identical tasks. If we observe an iteration, we generate a new task node of type iteration. The single child of this task node becomes the iterated event task.

Afterwards, we scan the ordered list of event task instances for instances of the iterated event task. We replace all instances

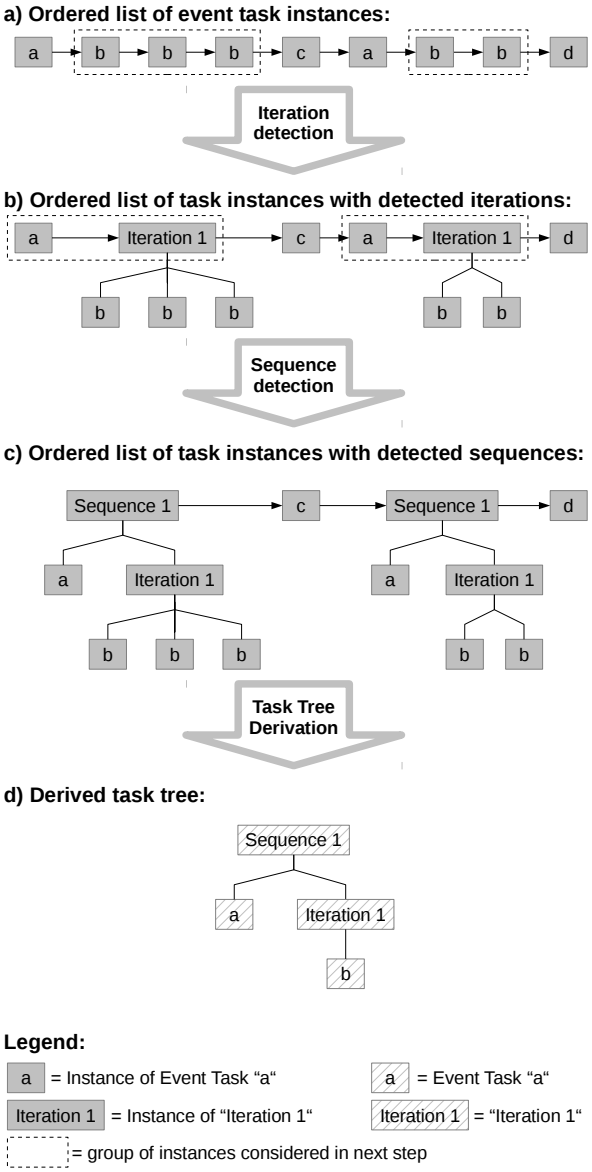


Figure 7. Example for the detection of iterations and sequences

of the iterated event task with instances of the new iteration. If the event task occurs only once, it is replaced by an iteration instance getting the replaced event task instance as its single child. If the event task occurs more than once subsequently, the subsequent instances are also replaced by a single iteration instance. This instance gets all replaced event task instances as its children. An example for such a replacement is shown in Figure 7a and 7b where Figure 7a depicts the event task instance list before the replacement and Figure 7b shows the event task instance list after the replacement. There, Event Task *b* is iterated several times (denoted by the dotted boxes in Figure 7a. We replace these instances in the ordered list with iteration instances having the replaced event task instances as their children.

4) Sequence Detection: After the iteration detection, we scan the list of task instances for multiple occurrences of the same subsequences. For the subsequence occurring most often and which is, therefore, most likely an occurrence of a logical subtask, we generate a new task node of type sequence. Its children are the tasks belonging to the identified subsequence. Each occurrence of the identified subsequence in the task instance list is replaced with an instance of the new sequence. Each instance gets as children the task instances to be replaced by the sequence instance. An example is shown in Figure 7 where Figure 7b depicts the task instance list before the replacement and Figure 7c shows the task instance list after the replacement. There, the subsequence of Event Task *a* and Iteration *1* occurs most often (two times) and is, therefore, replaced through instances of a sequence representing this subsequence.

The subsequences replaced through the sequence detection can have any length. At the minimum, they have a length of two. Our algorithm searches for the longest subsequences occurring most often and replaces it accordingly. If several subsequences have the same maximum occurrence count, we replace only the longest one. If several subsequences have the same maximum count and the same maximum length, we replace only the subsequence occurring first in the ordered list.

5) Alternating Repetition of Detections: The iteration and sequence detection on the ordered list of task instances are repeated alternately until no more replacements are possible. This is also visualized in Figure 5. Each time an iteration detection is done, all iterations are detected and replaced. This also includes iterations of detected sequences. For each sequence detection the longest sequence occurring most often is replaced. A detected sequence may include already detected sequences and iterations. For example, in Figure 7c the detected sequence contains a previously detected iteration.

In each alternating repetition of the iteration and sequence detection, the ordered list of task instances becomes shorter. This is because several task instances in the list are replaced by single sequence or iteration instances. But the replaced task instances as well as their order are preserved by making them children of their respective replacement. Hence, no details of the recorded events are lost.

If no more iterations or sequences are detected, the algorithm stops as shown in Figure 5. The resulting task instance list contains instances of detected task trees as well as event task instances that were neither iterated nor part of a sequence occurring more than once. Based on the instances of the detected task trees, we can derive the raw task tree. For the example in Figure 7a-c, the detected task tree is shown in Figure 7d. The detected task trees represent the lexical, syntactical and semantic level of design. The more recorded events are processed, the more complex and deeper task trees are created.

Within a recording of only one user session, specific subsequences occur only once. An example is the login process, which is usually done only at the beginning of a recorded user session. With our approach, such regularly occurring subsequences would not be detected if only one session was considered. Therefore, we consider several sessions of

different users at once for counting the number of occurrences of subsequences. Due to this, we also detect subsequences occurring seldom in individual sessions but often with respect to all recorded users of the website.

6) Characteristics of Detected Tasks: The task trees created so far consist only of sequences and iterations. Each detected task is characterized by several aspects with respect to the recorded events. First, each task t is associated a set of recorded events $r(t)$ based on which its structure was generated. For example, in Figure 7, the task *Iteration 1* is generated based on all instances of Event Task *b* and, hence, based on the corresponding events. Second, there is a function *depth*(t) that returns the depth of a task t. The depth is the number of levels a task has where the task itself is the first level, its children are the second, its grandchildren are the third, etc. The last level contains only event tasks and hence no further children exist. For example, the depth of task *Sequence 1* in Figure 7c is 3. Finally, it is possible to determine all instances $i(t)$ of a task t as they were generated during the task detection process.

E. Merging of Similar Sequences

The tasks detected through the alternating iteration and sequence detection follow strict structures and contain only iterations and sequences. Due to this strictness, two distinct tasks can be similar to each other and may describe the same overall task being executed in two slightly different variants. A typical example is the login process where some users use the tab key and some users use a mouse click to navigate from the user name field to the password field. A simplified example of two similar tasks is shown in Figure 9a. Both tasks start with Event Task a, have intermediate executions of Event Tasks c and e, and end with Event Task h.

To reduce the number of similar tasks, we perform the forth and final major step of our approach for generating task trees based on recorded user actions that is the merging of similar sequences (see Figure 1). This step also consists of several substeps: detection of similar sequences, determination of sequences to merge, adaptation of flattened task instances, iteration detection, and sequence detection. These substeps may be repeated several times depending, e.g., on how many similar sequences are detected. The order of execution of these substeps is shown in Figure 8. For each of these substeps and their repetition, we provide respective details in the following sections.

1) Detection of Similar Sequences: The first substep to merge similar sequences is to detect them (see substep one in Figure 8). For this, we compare all sequences with each other and calculate a similarity metric for each pair. For this, we first generate for each sequence t an ordered list $l(t)$ of event tasks covered by the sequence. This list contains the event tasks in the order they would be executed, if the sequence was executed with a minimum of event tasks, i.e., with all iterations being executed only once. This list is similar to the smallest flattened instance of the sequence with the distinction that it contains event tasks and not their respective instances. We create $l(t)$ by performing a depth first traversal of the structure of t and storing all leave nodes in a list in the order they were visited. The two lists of event tasks belonging to the sequences in Figure 9a are shown in Figure 9b.

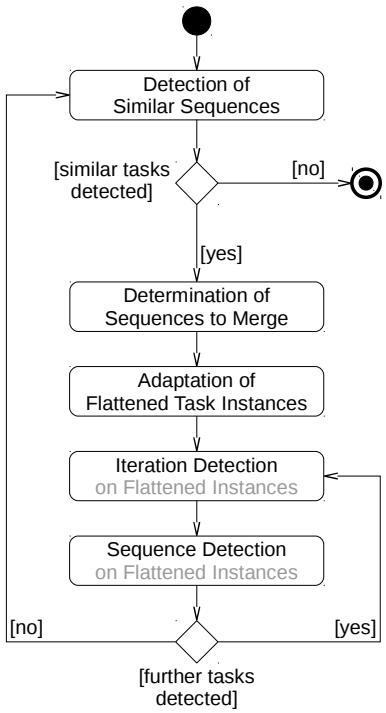


Figure 8. Basic process for merging similar sequences

In the next step, we compare these lists of event tasks with each other using Myers diff algorithm [11], which we adapted to compare lists of event tasks instead of strings. This diff can be seen as a function $D(l(t_1), l(t_2))$ that calculates a list of n deltas $d_1 \dots d_n$ between the two lists of event tasks determined for t_1 and t_2. The types of deltas that are detected are:

- *insert*: one or more subsequent event tasks occur only in the second list at a specified position;
- *delete*: one or more subsequent event tasks occur only in the first list at a specific position; and
- *change*: one or more subsequent event tasks in the first list are replaced by one or more subsequent event tasks in the other list.

For each delta d, the number of event tasks making up the delta is defined as $e(d)$. The three deltas determined for the sequences in Figure 9a are shown in Figure 9b. Based on the deltas, we calculate the similarity metric $s(t_1, t_2)$ of two sequences t_1 and t_2. This metric is calculated as the number of event tasks belonging to the determined deltas divided by the number of all event tasks of both sequences:

$$s(t_1, t_2) = \frac{\sum_{i=1}^{n} e(d_i)}{|l(t_1)| + |l(t_2)|} \quad (1)$$

In this calculation, we have to do a special consideration for scrolls. If an event task is a scroll, it must always be considered

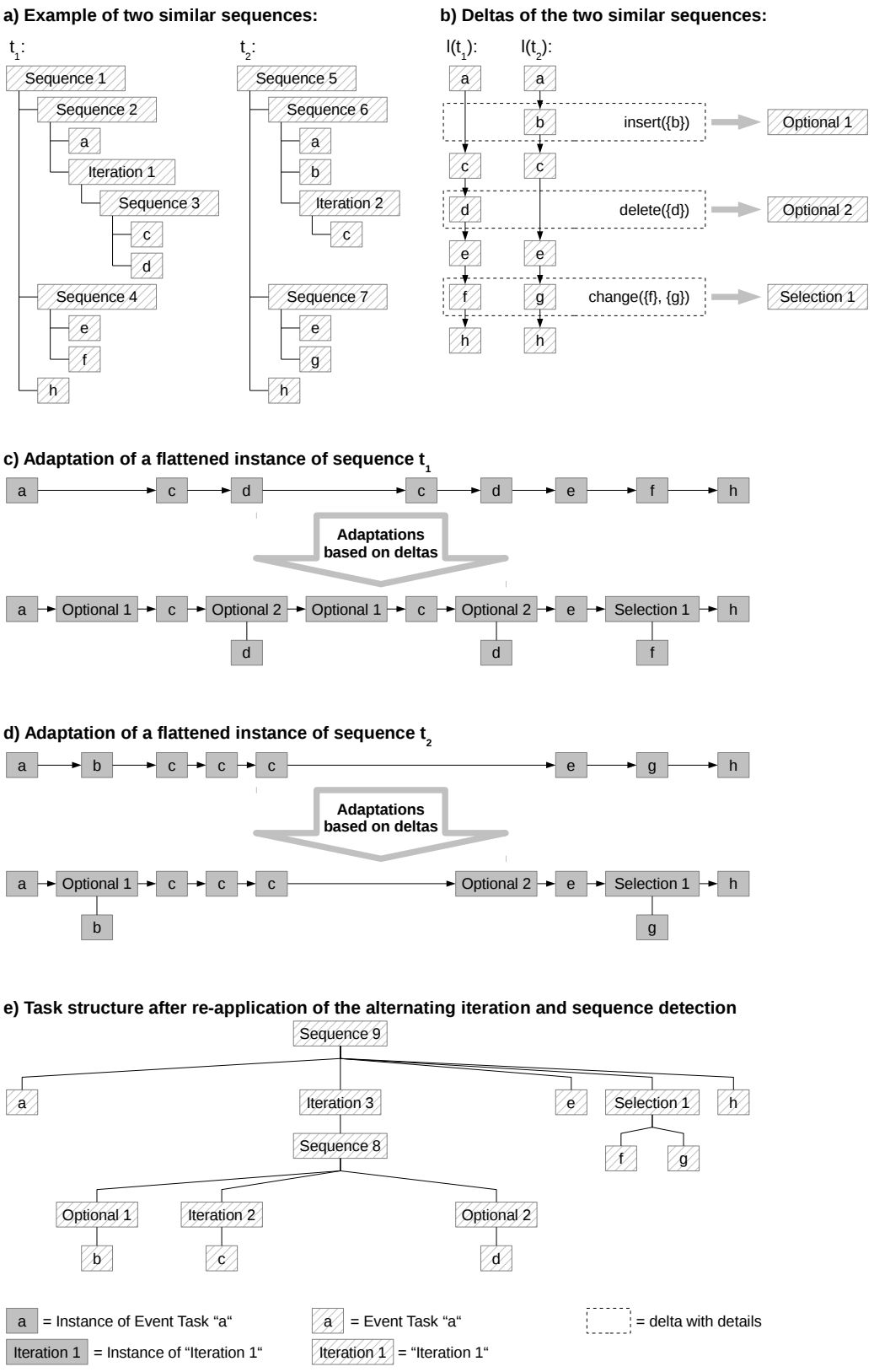


Figure 9. Example of two distinct but similar tasks generated by the alternating iteration and sequence detection and the process for merging them

distinct during the calculation of $s(t_1, t_2)$ independent of belonging to a delta or not. The reason for this is that scrolls do not have a semantic meaning in fulfilling a task. But two similar tasks should only be considered similar if their semantic meaning is similar. This should not be influenced by a vast occurrence of semantically unimportant event tasks like scrolling. Hence, we adapt the calculation of $s(t_1, t_2)$ accordingly. Let $o(d)$ be the number of scroll event tasks belonging to delta d and let $o(t_1, t_2))$ be the number of scroll event tasks belonging to $l(t_1)$ and $l(t_2)$. Through the following calculation of $s(t_1, t_2)$, we ensure that scrolls are not considered:

$$s(t_1, t_2) = \frac{o(t_1, t_2) + \sum_{i=1}^{n} (e(d_i) - o(d_i))}{|l(t_1)| + |l(t_2)|} \quad (2)$$

Through this calculation, $s(t_1, t_2)$ is smaller the more similar the sequences t_1 and t_2 are. If $l(t_1)$ is identical to $l(t_2)$ and if both lists do not contain any scrolls, $s(t_1, t_2)$ evaluates to 0 indicating the highest possible similarity between two sequences. However, as $s(t_1, t_2)$ does not consider subtask structures, the sequences may still be different with respect to execution order.

2) Determination of Sequences to Merge: After we calculated $s(t_1, t_2)$ for each pair of sequences p, we determine those pairs that have to be merged. This is the second substep for merging similar sequences as shown in Figure 8. Merging is only useful for sequences having a minimum level of similarity, i.e., $s(t_1, t_2)$ should be below or equal to a specific border that we call s_{max}. For merging, we determine only those pairs

1) for which $s(t_1, t_2) \leq s_{max}$
2) whose deltas are neither at the beginning nor at the end of $l(t_1)$ and $l(t_2)$ to ensure that deltas are always considered in their entirety
3) for which $s(t_i, t_j)$ is minimal and that are, hence, the pairs of the most similar sequences
4) for which none of the sequences is a direct or indirect parent of any sequence of another pair

For the remaining set P of pairs, we determine the set of sequences $T = t_1 \ldots t_2$ that are referred to by more than one pair in P. If there are no such sequences ($T = \emptyset$), we merge all pairs in P. If there are such sequences, we perform a further filtering based on them and consider only pairs of which at least one task is in T further. This is required to ensure that in such cases the sequences in T are merged always in the same order. For each sequence $t \in T$, we merge only the pair referring to t where the following characteristics of the pair are maximized or minimized in the given order in comparison to all other pairs referring to t:

- $r(t_1) + r(t_2)$ is maximized, i.e., both sequences of the pair were generated on a maximum of recorded events;
- $|i(t_1)| + |i(t_2)|$ is maximized, i.e., both sequences of the pair were most often executed;
- $\mathit{depth}(t_1) + \mathit{depth}(t_2)$ is minimized, i.e., both sequences of the pair have the smallest depth;
- the sum of the instances of both sequences as well as all their subtasks is maximized (subtasks are also executed in other task contexts and, hence, their number of instances is typically higher than the number of instances of a parent task);

If there are several pairs referring to a $t \in T$ for which these characteristics are maximized or minimized, we throw an exception and break up.

3) Adaptation of Flattened Task Instances: The result of the determination of similar sequences is a list of independent pairs of similar sequences. In the third substep of merging similar sequences (see Figure 8), we perform a merge for each of these pairs. The merging of a pair is done based on the flattened instances of both sequences t_1 and t_2 and on the deltas $D(l(t_1), l(t_2))$ determined for the pair. For this, the flattened instances of both sequences are created. Then, they are adapted based on the deltas to include instances of optionals and selections to reflect the deltas. Basically, we

- create instances of optionals to reflect insert and delete deltas as well as
- create instances of selections to reflect change deltas

Insert and delete deltas are handled in a similar fashion as they reflect the same situation: one or more event tasks are included in $l(t)$ of one sequence but missing in $l(t)$ of the other sequence. Let t_1 be a sequence where $l(t_1)$ includes the event tasks denoted by an insert or delete delta d and let t_2 be the sequence whose $l(t_2)$ does not include these event tasks. For the delta d, we generate an optional to reflect the delta. If the delta denotes exactly one event task, this event task becomes the single child of the optional. If the delta denotes more than one event tasks, we first generate a new intermediate sequence having the event tasks of d as children and set this sequence as the single child of the optional.

Afterwards, we adapt the flattened instances of both similar sequences. The flattened instances of sequence t_1 are adapted by replacing the instances of the event tasks denoted by d with instances of the new optional. If only one event task instance is replaced, this becomes the single child of the replacing optional instance. If more than one event task instances are replaced at once, we ensure that the replacing optional instance matches the task structure of the corresponding optional including intermediate sequences if any. The flattened instances of t_2 are adapted by integrating instances of the new optional at the position where the event tasks denoted by d are missing. These optional instances do not have children what reflects that the child task of the optional was not performed.

An example for handling an insert and a delete delta, both referring only to one event task, is shown in Figures 9b-d. In Figure 9b, we show the determined deltas and also name the two optionals that will be created to reflect them (Optional *1* for the insert and Optional *2* for the delete delta). Figures 9c and d show, how two exemplifying flattened instances of the similar sequences are adapted by integrating empty optional instances or replacing event task instances. For example, in Figure 9c, the instance of sequence t_1 does not contain an instance of Event Task *b* (indicated also by the insert delta). Hence, we integrate an empty instance of Optional *1* at the respective position between the instances of Event Task *a* and *c*. In addition, we replace the occurrence of Event Task *b* in the instance of t_2 (Figure 9d) with an instance of Optional *1* having the instance of *b* as its child.

To handle a change delta, we generate a selection. This selection gets two children, both representing the appropriate variants defined by the delta. If a variant consists of more than one event task, we again work with intermediate sequences to be able to reflect several subsequent event tasks at once. Afterwards, we adapt all flattened instances of both sequences. We replace each occurrence of the event tasks denoted by the delta with instances of the new selection. The children of the selection instances are set to reflect the executed variant. We again ensure that intermediate sequences, if any, are also correctly reflected in the selection instances. Figures 9b-9d show an example of this approach. Figure 9b shows a change delta for which we generate Selection *1*. Afterwards, we adapt the flattened example instances of t_1 and t_2 as shown in the Figures 9c and 9d. The event task instances *f* in the instance of t_1 and *g* in the instance of t_2 are both replaced by instances of Selection *1*. In both cases, the selection instance has the replaced event task instance, i.e., the selected variant, as its single child.

4) Iteration and Sequence Detection: After all flattened instances of a sequence pair to be merged are adapted according to the above rules, we reapply our sequence and iteration detection on all adapted flattened instances of both sequences (see substeps four and five in Figure 8). Both substeps are repeated until no more iterations or sequences are detected in the flattened instances. Afterwards, we get a new task structure as replacement for both sequences that now includes selections or optionals to reflect the different task variants. The result of the reapplication of our approach on the adapted flattened instances of Figures 9c and 9d is shown in Figure 9e.

The reapplication of the sequence and iteration detection is always done on at least two flattened task instances (one for each sequence of the merged pair). Furthermore, the flattened instances are adapted in a way so that their structure is the same. Hence, in the last cycle of the sequence and iteration detection, we automatically detect a sequence covering the whole new structure. This new sequence becomes the replacement for both merged sequences.

5) Handling of Interleaving Iterations: Although two sequences are similar, they may differ in the possible iterations of event tasks. A typical example is shown in Figure 10. This figure displays two variants of task trees of a login process as they are generated by the alternating iteration and sequence detection of our approach. For navigating from the user name to the password field, the first variant includes the usage of a mouse click on the password field whereas the second variant uses the pressing of the tabulator key. The merging process described in the preceding paragraphs would consider both sequences as similar as they differ only at two of ten event tasks and would merge them. But here, merging must be prevented. The reason is, that the behavior of a click on the password field and the usage of the tabulator key is different with respect to the focus state of the GUI. A repeated click on the password field leaves the focus on the password field. But a repeated usage of the tabulator key will move the focus always to the next element of a form. Hence, depending on the event tasks in both variants of the example the allowed iterations in the variants differ. We call this situation *interleaving iterations*.

Interleaving iterations cannot be merged by our approach. To handle such similar sequences anyway, we determine interleaving iterations before merging. In the example in Figure 10, these are the iterations marked with the grey arrows. If we find interleaving iterations, we change our merging process. First, we calculate a variant of $l(t)$ that we call $l'(t)$. This variant results from a depth first traversal of task t where event tasks are added to $l'(t)$ and interleaving iterations are not traversed but also directly added to $l'(t)$. As a result, $l'(t)$ contains either interleaving iterations or event tasks that are not part of an interleaving iteration. We then apply the diff algorithm $D(l'(t_1), l'(t_2))$ to get an adapted set of deltas. Second, when adapting the flattened instances of two similar sequences, we do not consider fully flattened instances but prevent flattening instances of the interleaving iterations. Furthermore, we consider only the adapted set of deltas when creating optionals and selections. The remaining steps stay the same. Through this, we adapt our overall process to preserve interleaving iterations and, hence, to be correct with respect to allowed or possible repetitions of actions.

6) Reapplication of Similar Sequence Merging: As shown in Figure 8, the detection and merging of similar sequences is repeated until no more similar sequences are found whose similarity level is low enough. This also includes merging of sequences being themselves results of merging or that have results of merging as their subtasks. Due to this, two compared and potentially merged tasks may include selections and optionals that need special considerations when comparing and merging tasks.

When creating $l(t)$ of a task including an optional, the optional is traversed normally. Hence, $s(t_1, t_2)$ of two tasks that may contain optionals in addition to sequences and iterations is calculated the same way. We make some additional considerations when performing a merging of two task t_1 and t_2. It may be the case, that elements of $l(t_1)$ being equal in $l(t_2)$ may be optional in t_1 but not in t_2 because of a parent optional. An example is shown in Figure 11a. There Event Task c is optional in task t_1 but mandatory in t_2. But this does not show up as a delta between $l(t_1)$ and $l(t_2)$ as seen in Figure 11b. To anyway preserve such optionals, we first identify those elements of $l(t_1)$ that have an optional as parent. Instances of these optionals are not flattened when creating the flattened instances of t_1. Afterwards, we ensure that all elements of $l(t_2)$ have an optional as parent if the corresponding elements of $l(t_1)$ also have an optional as its parent. If required, an optional is introduced. When creating the flattened instances of t_2, instances of these optionals also remain unflattened. As a result, all flattened instances of both tasks t_1 and t_2 will contain event task instances and optional instances. This ensures that the optionals contained in t_1 are preserved. Figures 11c and 11d show how the instances of the sequences t_1 and t_2 are flattened by preserving (Figure 11c) and respectively introducing (Figure 11d) an optional for Event Task c.

Selections require a more complex handling. If a task contains a selection, this selection can not be traversed when creating $l(t)$. Hence, $l(t)$ of a task containing a selection includes event tasks and selections. An example for this is also

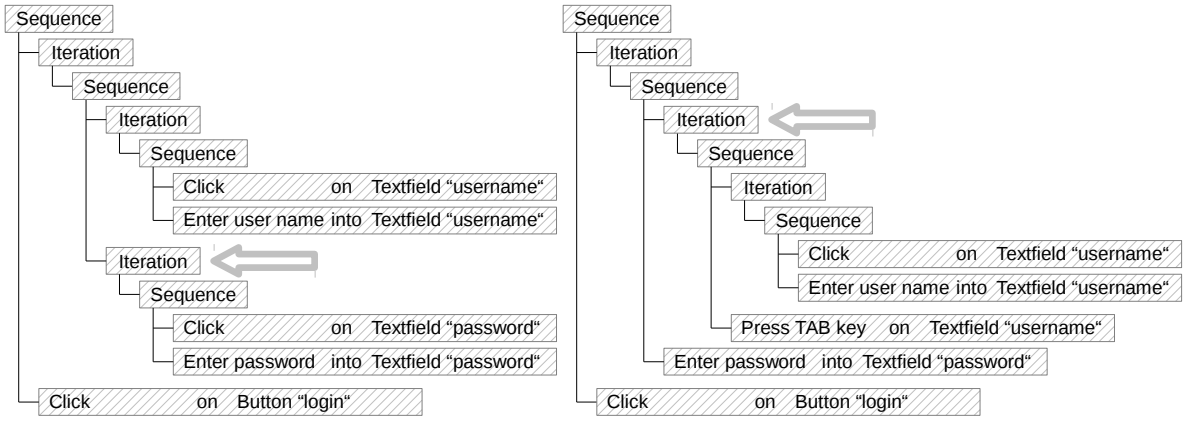


Figure 10. Example of two similar tasks with interleaving iterations

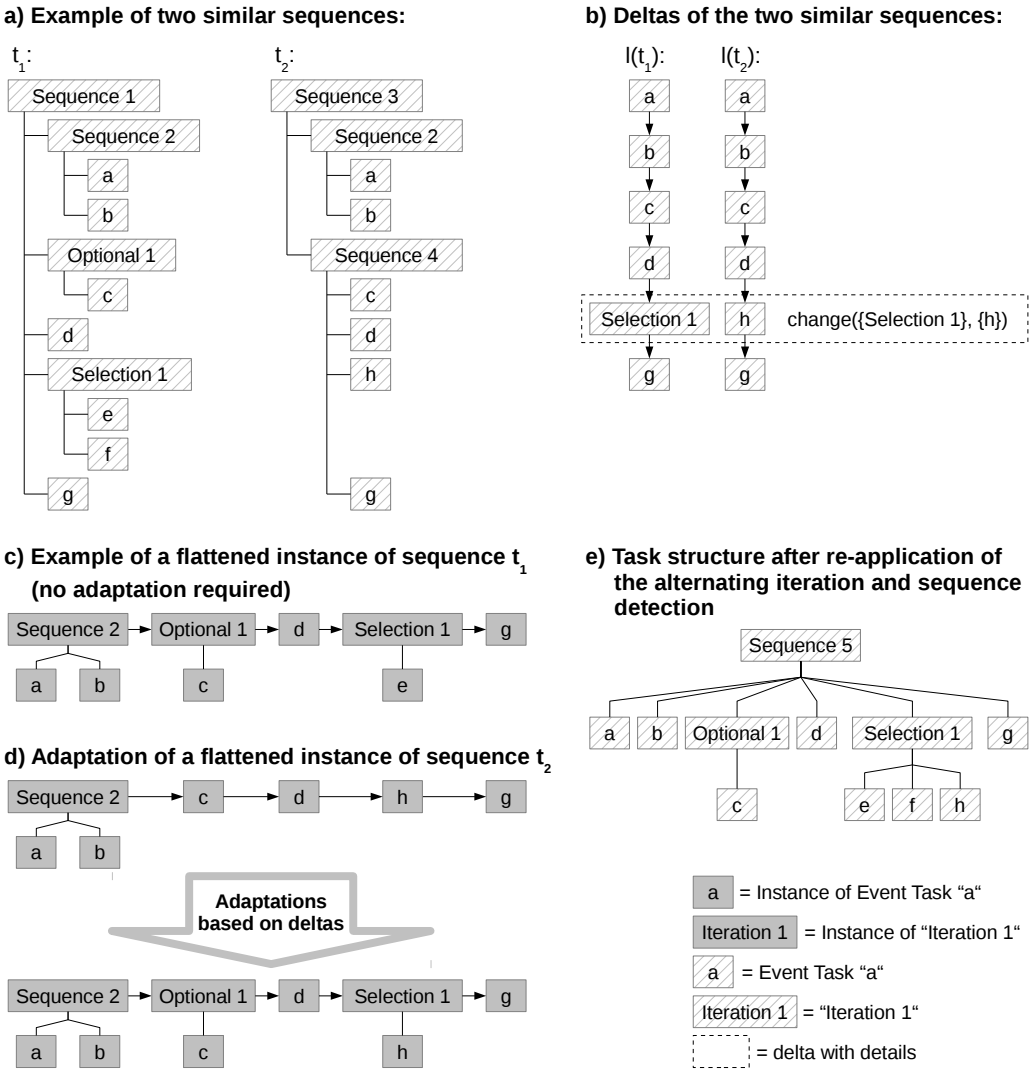


Figure 11. Example of merging two similar tasks containing optionals and selections

shown in Figure 11. There, Sequence t_1 contains a selection that is not flattened and hence shows up in $l(t_1)$.

Although a selection is not flattened when creating $l(t)$, it may represent several event tasks at once. Thus, we adapt the calculation of $s(t_1, t_2)$ to still correctly represent the ratio of different event tasks for the comparison of t_1 and t_2 even if one of them contains a selection. For this, we calculate the average number of event tasks $e(p)$ covered by each selection $p \in l(t_1) \cup l(t_2)$. This is done by creating $l(c)$ for each child c of a selection and then calculating the average length of these lists. Furthermore, if a delta d includes one or more selections, then $e'(d)$ is the sum of the average event tasks covered by the selections and the remaining number of event tasks belonging to the delta:

$$e'(d) = \sum_{p \in d} e(p) + e(d) \tag{3}$$

Finally, the calculation of $s(t_1, t_2)$ is adapted to respect the average number of event tasks covered by selections as follows:

$$s(t_1, t_2) = \frac{o(t_1, t_2) + \sum_{i=1}^{n} (e'(d_i) - o(d_i))}{|l(t_1)| + |l(t_2)| + \sum_{p \in l(t_1) \cup l(t_2)} (e(p) - 1)} \tag{4}$$

For merging two sequences, where one contains one or more selections, we do not flatten selection instances. This is to preserve the selection instance and its selected variant. Furthermore, if a selection spans a change delta completely, we adapt it to include the new variant in its children. An example for this is shown in Figure 11a. There, t_1 includes a selection. When flattening an instance of t_1 (Figure 11c) the instance of the selection is preserved and not flattened. When adapting the flattened instance of the similar task t_2, the selection is extended with a further variant (Event Task h) as the delta between t_1 and t_2 is fully spanned by the selection. Hence, the flattened instance of t_2 is also adapted by replacing the instance of Event Task h with an instance of the selection having h as the selected variant.

A further consideration in our merging process is the preservation of subtasks being either fully covered by a delta or representing common parts of two similar tasks. An example is shown in Figure 11a. Both sequences start with the sub task Sequence 2. When merging t_1 and t_2, this subtask and all its instances can and should be preserved. For this, when flattening the instances of two similar tasks, we do not flatten instances of subtasks being either common for both tasks or being fully covered by a delta. In the example in Figure 11c and d, the instances of Sequence 2 stay unflattened.

For each detected optional and selection as well as for intermediate sequences, we ensure that the same task is created only once. For example, if during several merges an optional of a specific event task must be created, the optional is created only during the first merge and then reused in the other merges. This also requires to detect if a generated task matches the requirements of a new task to be created.

F. Usability Evaluation

We utilize the generated task trees for automated usability evaluations. For this, we consider violations of generally accepted usability heuristics (e.g., as provided in [12]) and define patterns for their reflection in task trees. We then filter our task trees for these patterns and reason on potential usability defects. This is possible, as the generated task trees represent effective user behavior. For example, we analyze typical action combinations or navigation patterns users perform on a website and detect if their efficiency could be improved. The details of this approach can be found in [6].

III. Proof-of-concept Implementation

To show that our method is feasible, we implemented it based on the tool suite for Automatic Quality Engineering of Event-driven Software (AutoQUEST) [13]. The AutoQUEST platform provides diverse methods for assessing the quality of software. AutoQUEST's internal algorithms operate on abstract events, which makes AutoQUEST independent of the platform of an assessed software. AutoQUEST's modular architecture allows the extension with modules to support algorithms for quality assurance, as well as feeding AutoQUEST with events of a yet unsupported software platform. In the following, we describe how we utilized and extended AutoQUEST to implement our method.

A. User Interaction Tracing

AutoQUEST provides basic functionality for tracing user actions on different platforms including websites. For this, it uses techniques from GUI testing, e.g., for capture/replay testing [14]. For monitoring a website only a JavaScript needs to be added to each of the pages of the website. In modern content management systems, this can be configured centrally and easily. The JavaScript is served by a monitoring server shipped with AutoQUEST. It automatically records events caused by user actions. After a specific amount of events is recorded, or if the user switches the page, the script sends the events to the AutoQUEST server that stores them into log files. Using a dedicated parser, these log files can then be fed into AutoQUEST for further processing.

An excerpt of a trace of AutoQUESTs website monitor showing a mouse click and a scroll event on a web page is shown in Figure 12. Both events denote their respective type, a timestamp, and meta information like the coordinates in the click event. Furthermore, both events refer to a target, i.e., the element of the webpage, on which the event was observed. The identifiers of the targets can be resolved through other information stored in the log file, as well.

B. Task Tree Generation

For our proof of concept implementation, we extended AutoQUEST with capabilities to generate task trees based on traces. The implementation follows the overall process described in Section II. The implementations of the data preparation and the iteration detection are straightforward and, therefore, not described in more detail.

```
<event type="onclick">
 <param name="X" value="87"/>
 <param name="Y" value="213"/>
 <param name="target" value="id1"/>
 <param name="timestamp" value="1375177632056"/>
</event>
<event type="onscroll">
 <param name="scrollX" value="-1"/>
 <param name="scrollY" value="-1"/>
 <param name="target" value="id2"/>
 <param name="timestamp" value="1375177632900"/>
</event>
```

Figure 12. Example for a trace recorded with AutoQUEST's HTML monitor

1) Sequence Detection Implementation: For identifying and counting subsequences occurring several times, we reused and extended a data structure provided with AutoQUEST called trie [8]. A trie in AutoQUEST is a tree structure used for representing occurrences of subsequences in a sequence. In our case, we use the trie for representing subsequences of tasks in the ordered list of tasks considered for the next sequence detection. An example for a trie is shown in Figure 13.

Each node in a trie represents a task subsequence. The length of the represented subsequence is equal to the distance of the node to the root node of the trie. The root node of the trie represents the empty subsequence. The children of the root node (in Figure 13 all nodes on Level 1) represent the subsequences of length 1 occurring in the trace, i.e., all different tasks. The grand children of the root node (in Figure 13 all nodes on Level 2) represent the subsequences of length two as their distance to the root node is two, etc. The subsequence represented by a node can be determined by following the path through the trie starting from the root node and ending at the respective node. The length of the longest subsequence represented through a node in the trie is defined as the depth of the trie. The depth of the trie in Figure 13 is three.

Each node in a trie is assigned a counter. This counter defines the number of occurrences of the subsequence represented by the node. The counter of the root node is ignored. The example trie in Figure 13 represents the event tasks for the trace of Figure 4. The trie shows that the event task of clicking on the user name text field occurs twice and that both times it is succeeded by entering some text, i.e., a user name, into the text field. The event of clicking the login button is not succeeded by any other event task.

We calculate a trie each time a sequence detection on the ordered list of tasks is done. Based on the trie, we are able to identify the longest subsequence of tasks with a minimal length of two occurring most often. The number of occurrences is determined through the counts assigned to each node in the trie. The length of the subsequence is determined by the distance of the trie node representing the most occurring subsequence to the root node of the trie.

If the length of the identified subsequence is identical to the depth of the trie, we cannot decide if there is a longer subsequence with the same count. We, therefore, increase the depth of the trie until the depth is larger than the length of the longest subsequence occurring most often. In Figure 13, the longest subsequence occurring most often is clicking on the user name text field and entering a user name. This subsequence occurs twice and there is no other subsequence of the same or a longer length occurring more often. Therefore, all occurrences of this subsequence in the ordered list of tasks is replaced through a task node of type sequence.

2) Comparison of Tasks: An important challenge in our implementation was the comparison of tasks. Tasks need to be compared very often either for compiling the trie or for detecting iterations. For an effective task generation, some tasks must be considered equal although they are different. An example is a task and an iteration of this task. Both must be considered identical if the iteration is executed only once. Another example is shown in Figure 13. The represented trie contains nodes for the event tasks representing the entering of text into the user name text field. Although different text is entered in the respective events, the respective event tasks need to be considered identical for a correct trie calculation. Therefore, we implemented a mechanism to be able to perform complex task comparisons. In addition to other comparisons, it is able to compare a task A with an iteration of a task B and considers them as equal if task A is equal to task B.

3) Merging of Similar Sequences: Due to the large number of sequence comparisons required for detecting similar sequences, we implemented the comparisons to be executed in parallel. For this, we schedule several threads each performing a bucket of all required comparisons. The number of threads executed in parallel can be configured and should match the number of available cores on a machine. Each thread searches for those sequence pairs in its buckets that have the lowest similarity level and returns them. Afterwards, the results of all threads are joined and again the most similar sequences are filtered out.

Furthermore, we implemented some checks to ensure that the merging was correct. For example, we ensure that the original flattened instances of two merged tasks are identical to the flattened instances of their replacement task. Additionally, we performed random manual checks of what is merged and if the merge results are correct.

IV. Case Studies

For the validation of our approach, we performed two case studies. With these case studies, we intended to show that our approach is feasible and able to generate task trees based on recorded user actions. For this, we first recorded user actions on two websites. Then we used our approach, i.e., its implementation, to generate task trees. Finally, we evaluated the correctness of the detected task trees through manual inspection and comparison with the structure and available user actions of the website and its pages. However, due to the partially large number of generated tasks, this inspection was only done for a subset of tasks.

For the first case study, we traced the interaction of users of our research website [15]. This website provides three major information categories to its users: information about our research group members, information about our research and corresponding projects, as well as information about the

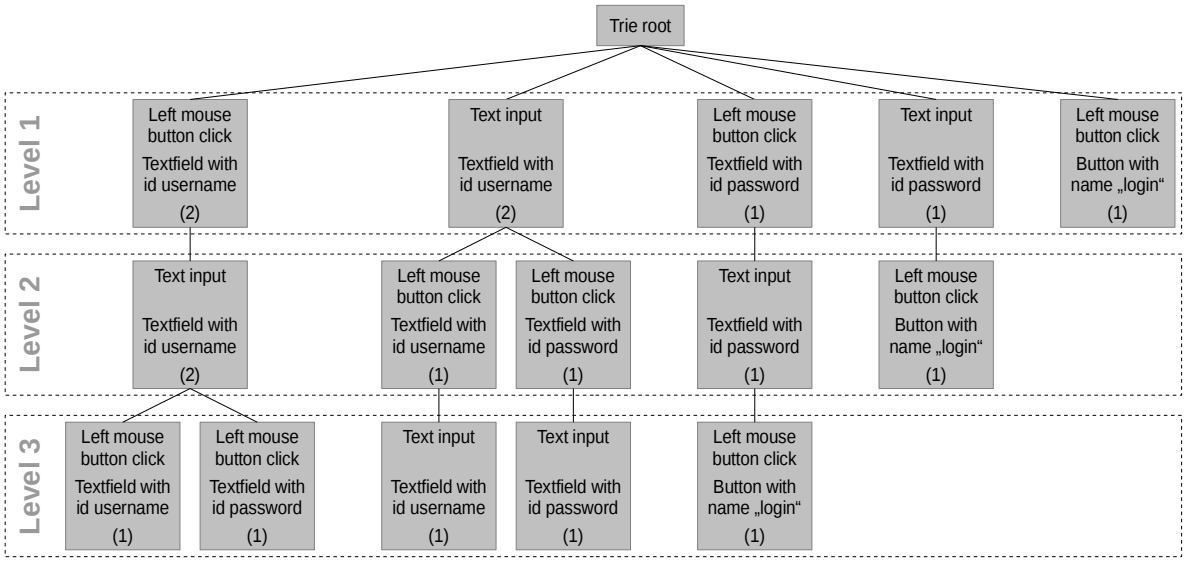


Figure 13. Trie generated based on the trace in Figure 4

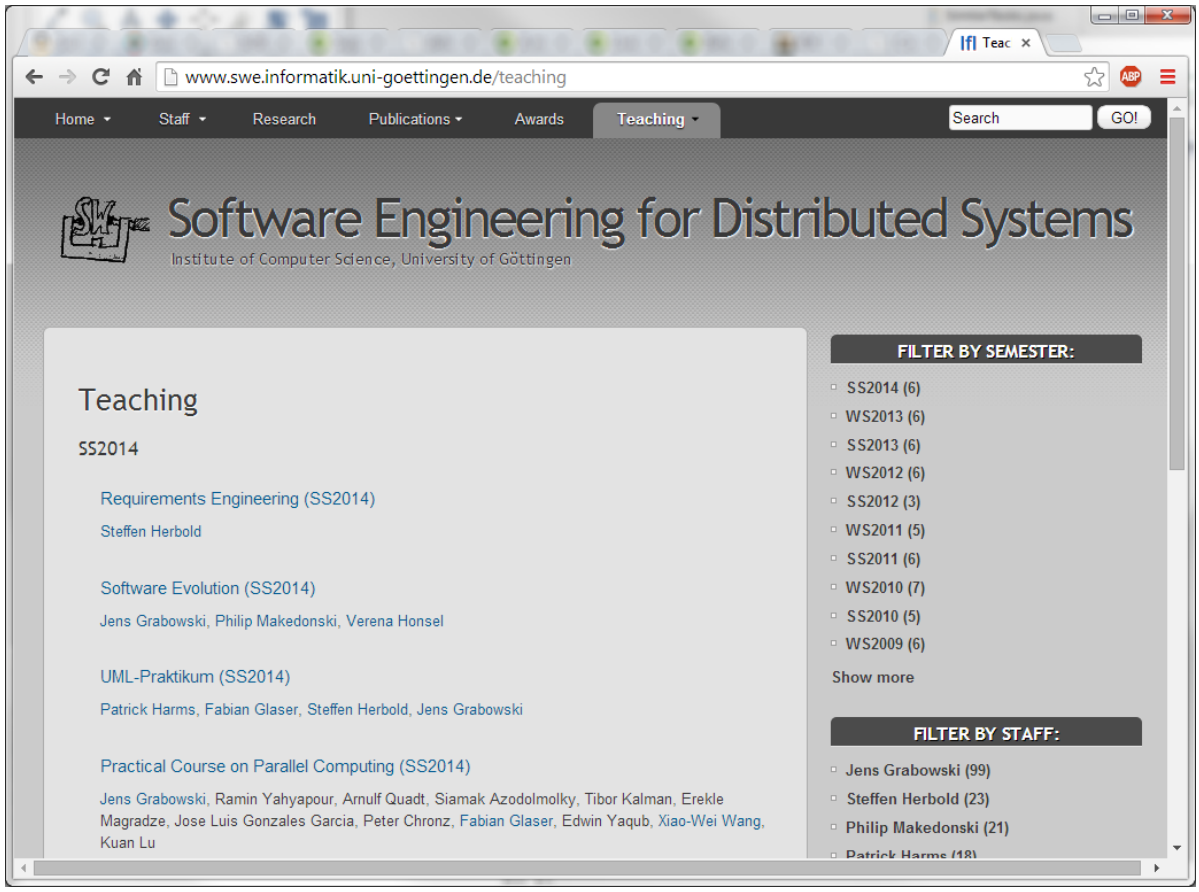


Figure 14. Screenshot of the Research Website of the first case study

courses we offer to students at our institute. A screenshot of the page listing the offered courses is shown in Figure 14.

We integrated the HTML monitor of AutoQUEST in our content management system. We then recorded interactions of

TABLE I. RECORDED AND CONSIDERED EVENTS IN THE TWO MAJOR CASE STUDIES

	Case Study 1 Website of Research Group	**Case Study 2** Application Portal
Start of Recording	30 June 2013	25 October 2013
End of Recording	4 March 2014	7 March 2014
Recorded Users	1,356	555
Recorded User Sessions	8,113	4,129
Considered User Sessions	6,587	3,635
Recorded Events	63,127	350,368
Relevant Events	38,070	306,568
Double Clicks	741	6,437
Focus Changes	9,950	89,825
Considered Events	27,379	210,306
Different Events	1,202	1,897
Generated Tasks Without Merging	1,847	10,634
Sequences	1,431	9,530
Iterations	416	1,104
Generated Tasks With Merging	1,842	10,663
Sequences	1,419	9,361
Iterations	416	1,133
Selections	3	81
Optionals	4	88

1,356 users over a period of 8 months. Afterwards, we fed the gathered traces containing more than 63,000 events into AutoQUEST and generated over 1,800 task trees based on this. Further details of the case study are listed in the middle column of Table I.

This case study showed that the task tree generation was feasible in general. The generated task trees represented user behavior occurring several times. As an example, several users opened the initial web page and navigated to our teaching page (shown in Figure 14). From there, they navigated to the information about a specific course by clicking on one of the respective links.

In the first case study, we set s_{max} to 0.33, i.e., we considered two sequences as similar if less than or equal to one third of the event task lists of both sequences are different (including scrolls). We chose this border to reflect that we expect at least twice as many commonalities than differences between sequences to consider them as similar. Using this parameter, the merging of tasks in the first case study found 12 similar tasks and merged them resulting in three selections and four optionals. One optional was detected and reused in 6 merges, one selection in two merges. The merged tasks were executed quite seldom and covered at most 262 events per similar task pair. The optional that was integrated most often represents an optionality of scrolling.

The first case study revealed that our mechanism must be careful with respect to privacy protection. Our research website includes a log-in mechanism for being able to change its content. The first version of the tracing mechanism also traced user names and passwords of all users that logged in on the website. As this was a severe security issue, we adapted the tracing mechanism to ignore password fields in general. Furthermore, a website can be instrumented in a way, so that contents of selected text fields, e.g., fields for entering a user name, are not traced anymore.

In our second case study, we traced the users of an application portal of our university over a period of 4 months. This case study traced over 550 users producing more than 350,000 events resulting in 10,663 generated task trees. The details for this case study are listed in the right column of Table I. As in the first case study, we set s_{max} to 0.33.

When feeding the data of the second case study into AutoQUEST, we initially observed performance problems of our approach. Especially, the large number of distinct events caused the creation of a large trie for sequence detection. We, therefore, implemented several optimizations. For example, click events on the same button but with different coordinates are now treated as the same event task. However, click events on other website elements are still considered different, if their coordinates differ. Furthermore, we made intensive use of the mapping of common page elements and thereby reduced the number of different event tasks. This was done as on some pages of the portal, tables were used to represent content of similar meaning (e.g., a list of the application data entered by a user). For each row in such a table, there were buttons to edit the content. Although each of these buttons refer to different content to be edited, the basic semantic of these buttons is considered the same: edit the content belonging to the row. Hence, we considered these buttons as common page elements.

The second case study showed that our approach is able to correctly identify effective user behavior. The application portal also provides a login mechanism. Our task tree generation created several different task trees for the login process of users. One of them showed the behavior of those users using the mouse to set the focus on the password field after having entered the user name. Another task tree showed the usage of the tabulator key instead. The merging process correctly identified these two tasks as similar and merged them considering the interleaving iterations. A visualization of the merge result as displayed by AutoQUEST is shown in Figure 15. The example shows, that many iterations are generated in the task tree. This is due to the fact, that some users corrected the entered data several times. Furthermore, if the users entered wrong credentials, the website returned to the same page and the users started the login process again. The selection resulting from the merge is very large due to the handling of interleaving iterations.

In the second case study, the merging process identified more similar task than in the first case study and performed 224 mergings. This resulted in 81 selections and 88 optionals. The selection that covered most recorded events was a selection between entering a date into five different text fields and was mainly integrated into tasks representing the usage of a date chooser. The optional that covered most recorded events was also here an optional of scrolling. Due to the merging, the overall number of tasks increased, as new selections and optionals were introduced but fewer other tasks (sequences) where discarded. Furthermore, more iterations were generated. This is because two similar tasks may have been executed subsequently. Through the merging, they were afterwards

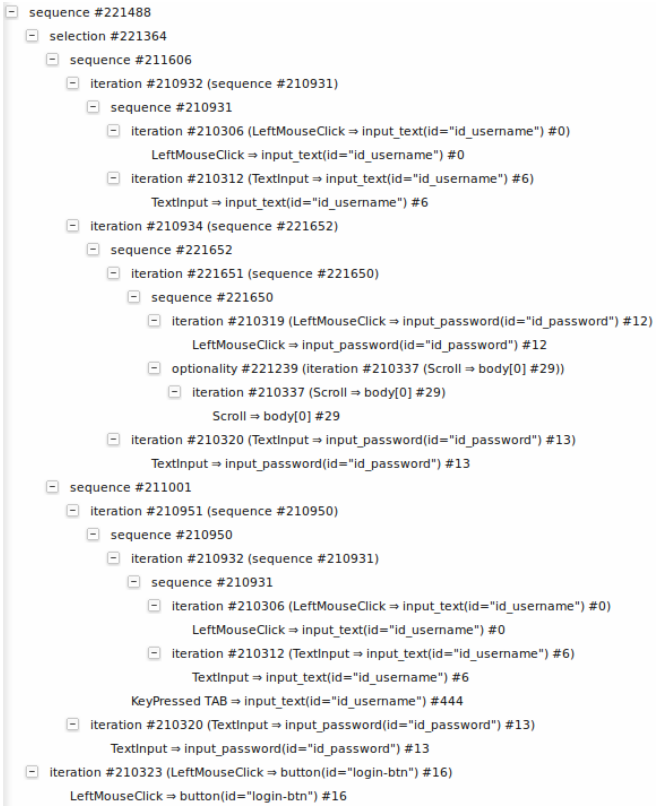


Figure 15. Task tree generated in the context of the second case study

considered the same task. Hence, their subsequent execution was correctly identified as an iteration of the merged task.

V. Discussion

The generated task trees represent the effective user behavior. This is important to analyze the usage of a monitored website, e.g., with respect to usability. Our approach is also able to identify distinct ways of executing semantically equal tasks and merge them into a single task.

At each repetition, the detection of sequences chooses the longest subsequence occurring most often and replaces it as described. This heuristic prefers shorter sequences as the count of a subsequence decreases with an increasing length. The resulting task trees are, therefore, deeply structured. Hence, it would be better to apply a more sophisticated heuristic such as selecting a subsequence occurring more seldom but being much longer.

Currently, we identify tasks that are executed only seldom. For example, we generate sequences for event combinations that happened only twice during user tracing. In the future, we plan to perform a filtering so that tasks must have a minimum amount of covered recorded events to be considered as tasks. This can be seen as a measure for the evidence of a task to be really a common task for all users.

For user interactions or tasks there may be pre or post conditions. For example, an iteration can be repeated a minimum or maximum amount of times. Our approach is not able to detect these conditions. Therefore, the task tree structures that we generate do not include notations for conditions.

The merging process allows to have in the end less tasks describing the semantically same task with slightly different execution variants. However, the merged tasks and their parent tasks may not be fully correct anymore with respect to the possible action sequences they represent. For example, in the second case study, the process merged several similar tasks of using a date chooser. All these tasks were identical except for the event task that finally entered the selected date into the respective text field. The merging process created a selection of these event tasks and merged several date chooser usages into a merged task. The last element of this merged task was a selection of entering a date into several available date fields on different pages of the website. Unfortunately, the merged task was a child of other parent tasks, e.g., of a parent task that was executed only on a specific page of the website. Hence, because the merged task also represented usages of the date chooser on other pages of the website, not all of its execution variants were valid in the context of this parent task. But for analyzing the usage of the date chooser, this merging was very helpful.

As a result, also invalid parent tasks were created through merging. This also applies for all parent hierarchies of a merged task. To check, if a parent task becomes invalid trough merging, a manual inspection was required. In the first case study, we did not observe this issue. In the second case study, we identified 8 of the 81 selections to cause invalid parent tasks. All were related to the date chooser usage. The cause for this issue is the consideration of common page elements. All date choosers on all different pages were considered as common page elements. But although being positioned at the same location in the DOMs of the different pages, they were semantically related to a specific page and a specific date field and, hence, should not be considered common. Therefore, this issue is not caused by the merging process itself but by the data preparation.

The merging process is based on the detection of similar tasks using Myers diff algorithm. In this work, we have not evaluated if the application of other diff algorithms would reveal other merging results. Hence, the results of the merging process may depend on the used diff algorithm. Furthermore, we set s_{max} in both case studies to a fixed value. We have not evaluated how a change on s_{max} may affect the merge result.

In some situations, our merging process does not yet fully correctly merge execution variants. For example, our approach created several selections, of which at least one child was a task and another child contained the task too as some embedded child. The simplest example is a selection of a task and an optional of the same task. Hence, some generated task structures are still more complex than they could be and need to be further refined.

VI. Related Work

In this section, we refer to related work. We start with similar work on recording user actions. Then, we consider

different approaches for task modeling and compare them with our work. Finally, we compare our approach with other attempts to generate task trees.

A. Recording of user actions

Nowadays, the idea of tracing user actions on websites is often applied in the context of web analytics, e.g., with Google Analytics [16] and Piwik [17]. Furthermore, there exist several tools that can be used to trace software based on other platforms than HTML. In contrast to our approach, the level of detail of the information recorded by existing tracing mechanisms varies and is often rather low. For example, Piwik does not record individual clicks.

To get more detailed recordings, other approaches, e.g, the one used by WebRemUsine [5], require Java applets or other mechanisms to store the recorded user interactions on the client and send them subsequently to the server. UsaProxy [18] provides an Hyper-Text Transfer Protocol (HTTP) proxy that is located between a web server and its clients. The proxy adapts each HTML document requested by a client and inserts a reference to a Javascript. This script automatically records detailed user actions at client side and sends them via Asynchronous JavaScript and XML (AJAX) [19] to the proxy that in turn stores them. Our approach for recording user actions on websites also utilizes an AJAX approach but without using a proxy. Instead, our Javascript is integrated in each website using mechanism of the content management system of a website. This has the advantage, that we can distinguish between monitored and unmonitored parts of a website. Furthermore, instead of using a standard way for instrumenting all pages of a website as done by UsaProxy, our approach allows to consider website specific technical challenges when adding the Javascript to the pages.

B. Task Modeling

Task models are used to describe the actions a person has to perform to reach a specified goal. In the context of website usage, they allow to define, which and how website elements are to be used to accomplish one of the tasks the website was developed for [5]. Task models usually cover task decomposition, task flow specification, object modeling, and task world modeling [20]. Our approach is restricted to task decomposition and task flow specification. Task models can be either used for aiding design, validating design decisions, or, in the most formal way, generating user interfaces [20]. Our task trees aid design and are restricted to summative validation.

Van Welie et al. [20] developed a common ontology for task models as a basis for a harmonized comparison of different task modeling approaches. The ontology covers concepts and their relationships that are typically used in task modeling. The concept that is covered by our approach is *Task* with its more concrete variants *basic task* and *user action* (identical to the term *action* in our approach). Van Welie et al. define a basic task as "... a task for which a system provides a single function. Usually[,] basic tasks are further decomposed into user actions and system operations". The task structures that we generate in our approach are mainly on the level of basic tasks. However, the root nodes of our generated task trees are similar to Van Welies unit tasks that they consider "... as the simplest task that a user really wants to perform". We do not cover other concepts of Van Welies ontology. The relationships defined in the ontology that are also covered by our approach are *subtask* and *triggers*. The subtask-relationship defines the child tasks or actions of a task. The trigger-relationship defines the order of task execution. The trigger-relationship can have three different types of which we cover only *NEXT* and *OR*. NEXT indicates, which task is executed next what is covered by our temporal relationship sequence. OR indicates that there is a choice between several next tasks. This is covered in our approach by selection (simple choice), iterations (repeat or go on with next task/action), and optional (perform or skip task/action). Due to this possible mapping to Van Welies ontology, our task trees should be transformable into other task tree notations that can be mapped to the ontology, as well. But this is scheduled for future work.

Task trees are one possible variant for modeling tasks. The concept of task trees is applied, e.g., in Goals, Operators, Methods, and Selection Rules (GOMS) [21], TaskMODL [22], and ConcurTaskTrees [23] [7]. If the nodes of task trees define the temporal relationships for their children, they have the drawback that intermediate nodes may be required to fully specify the task flow [20]. An alternative for task modeling are workflow representations that do not have this drawback but require a time axis. In our approach, we use the basic concept of task trees, but apply it in a simplified manner.

C. Task Tree Generation

There have been several attempts to generate task trees automatically. For example, the Convenient, Rapid, Interactive Tool for Integrating Quick Usability Evaluations (CRITIQUE) [24] creates GOMS models based on recorded traces. A similar approach is proposed by John et al. [25]. ReverseAllUIs [2] generates task trees based on models of the GUI. The resulting task trees represent all available interactions a user can perform. In contrast to our work, these approaches do not generate task trees that represent the effective behavior of the users, but only a simplified or complete task tree of a website.

A further attempt to identify reoccurring user behavior is programming by example. Here, user actions are recorded to determine reoccurring action sequences. The system then offers the user an automation of the identified action sequence. An example of this work can be found in [26]. These approaches only attempt to locally optimize the usability, whereas we adopt a global view on the system.

Generating task trees for user actions is similar to the inference of a grammar for a language. The user actions are the words of a language that the user "speaks" to the software. The task tree is the grammar defining the language structure. However, current approaches for grammatical inference require the identification of sentences of the language before the derivation of the grammar [27]. For example, during one user session a user may execute several tasks or interrupt a task execution. This would lead to several sentences following each

other or incomplete sentences in a user session. Hence, to apply grammatical inference, it would be required to mark those actions in a recorded user session that together form a sentence and to drop incomplete sentences. This would practically not be feasible in the case of a large set of recorded user sessions. Our approach is capable of handling large amounts of recorded actions without requiring a manual marking of correct task executions.

VII. Summary and Outlook

In this paper, we described a method for generating task trees based on tracing user interactions. First, sequences and iterations of user actions are identified. Then, semantically similar sequences are merged. We implemented this method for websites and performed two case studies to validate its feasibility.

In our future work, we will improve and extend the task tree generation. We especially focus on filtering tasks based on a minimum number of covered recorded events. We will ensure that parent tasks will not become invalid due to the extension with execution variants by merging child tasks. Furthermore, we plan to implement a better heuristic for detecting more intuitive subsequences, a flattening algorithm for reducing the complexity of the generated task trees, and an export of our task trees into a format used by other tools, e.g., into the format utilized by the ConcurTaskTrees Environment [28]. In addition, we improve the existing AutoQUEST plug-ins and implement plug-ins for further platforms, e.g., for operating systems with a focus on touch-based interaction. Finally, we improve the automated usability evaluation based on the generated task trees and perform comparisons of the usability evaluation results depending on if the evaluation is based on merged or unmerged tasks.

Acknowledgment

This work was partially done in the context of the project MIDAS (Model and Inference Driven - Automated testing of Services Architectures), funded by the European Commission, project number 318786.

References

[1] P. Harms, S. Herbold, and J. Grabowski, “Trace-based task tree generation,” in Proceedings of the ACHI 2014, The Seventh International Conference on Advances in Computer-Human Interactions. ThinkMind, 2014, pp. 337–342.

[2] R. Bandelloni, F. Paternò, and C. Santoro, “Engineering interactive systems,” J. Gulliksen, M. B. Harning, P. Palanque, G. C. Veer, and J. Wesson, Eds. Berlin, Heidelberg: Springer-Verlag, 2008, ch. Reverse Engineering Cross-Modal User Interfaces for Ubiquitous Environments, pp. 285–302.

[3] F. Paternò, “Model-based tools for pervasive usability,” Interacting with Computers, vol. 17, no. 3, 2005, pp. 291–315.

[4] L. Paganelli and F. Paternò, “Tools for remote usability evaluation of web applications through browser logs and task models,” Behavior Research Methods, vol. 35, 2003, pp. 369–378.

[5] F. Paternò, “Tools for remote web usability evaluation,” in HCI International 2003. Proceedings of the 10th International Conference on Human-Computer Interaction. Vol.1, vol. 1. Erlbaum, 2003, pp. 828–832.

[6] P. Harms and J. Grabowski, “Usage-based automatic detection of usability smells,” in Proceedings of the HCSE 2014, The Fifth International Conference on Human-Centered Software Engineering, 2014.

[7] F. Paternò, “ConcurTaskTrees : An engineered approach to model-based design of interactive systems,” The Handbook of Analysis for HumanComputer Interaction, 1999, pp. 1–18.

[8] S. Herbold, “Usage-based Testing of Event-driven Software,” Ph.D. dissertation, University Göttingen, June 2012 (electronically published on http://webdoc.sub.gwdg.de/diss/2012/herbold/ [retrieved: 1, 2014]), 2012.

[9] R. J. Jacob, “User interface,” in Encyclopedia of Computer Science, ser. Encyclopedia of Computer Science, A. Ralston, E. Reilly, and D. Hemmendinger, Eds. Nature Publishing Group London, 2000, pp. 1821–1826.

[10] J. Foley, Computer Graphics: Principles and Practice, ser. Systems Programming Series. Addison-Wesley, 1996.

[11] E. Myers, “Ano(nd) difference algorithm and its variations,” Algorithmica, vol. 1, no. 1-4, 1986, pp. 251–266. [Online]. Available: http://dx.doi.org/10.1007/BF01840446

[12] U.S. Department of Health & Human Services. Usability.gov - improving the user experience - guidelines. [Online]. Available: http://guidelines.usability.gov/ [retrieved: 11, 2014] (2013)

[13] S. Herbold and P. Harms, “AutoQUEST - Automated Quality Engineering of Event-driven Software,” March 2013.

[14] J. H. Hicinbothom and W. W. Zachary, “A Tool for Automatically Generating Transcripts of Human-Computer Interaction,” in Human Factors and Ergonomics Society 37th Annual Meeting, vol. 2 of Special Sessions, 1993, p. 1042.

[15] Software Engineering for Distributed Systems Group. Software Engineering for Distributed Systems. [Online]. Available: http://www.swe.informatik.uni-goettingen.de/ [retrieved: 7, 2014] (2014)

[16] Google. Google analytics. [Online]. Available: http://www.google.com/analytics/ [retrieved: 11, 2014] (2014)

[17] Piwik.org. Piwik - real time analytics reports for your websites. [Online]. Available: http://de.piwik.org/ [retrieved: 11, 2014] (2014)

[18] R. Atterer, “Usability tool support for model-based web development,” dissertation, Oktober 2008. [Online]. Available: http://nbn-resolving.de/urn:nbn:de:bvb:19-92963

[19] J. J. Garrett, “Ajax: A New Approach to Web Applications,” 2005, adaptive Path LLC, http://www.adaptivepath.com/publications/essays/archives/000385.php.

[20] M. Van Welie, G. C. Van Der Veer, and A. Eliëns, “An ontology for task world models,” in Proceedings of DSV-IS98, Abingdon, 1998. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.4415

[21] Q. Limbourg and J. Vanderdonckt, “Comparing task models for user interface design,” in The Handbook of Task Analysis for Human-Computer Interaction, D. Diaper and N. Stanton, Eds. Mahwah: Lawrence Erlbaum Associates, 2004.

[22] H. Trætteberg, Model-based user interface design. Information Systems Group, Department of Computer and Information Sciences, Faculty of Information Technology, Mathematics and Electrical Engineering, Norwegian University of Science and Technology, May 2002.

[23] F. Paternò, C. Mancini, and S. Meniconi, “ConcurTaskTrees: A diagrammatic notation for specifying task models,” in Proceedings of the IFIP TC13 International Conference on Human-Computer Interaction, ser. INTERACT ’97. London, UK, UK: Chapman & Hall, Ltd., 1997, pp. 362–369.

[24] S. E. Hudson, B. E. John, K. Knudsen, and M. D. Byrne, “A tool for creating predictive performance models from user interface demonstrations,” in Proceedings of the 12th annual ACM symposium on User interface software and technology, ser. UIST ’99. New York, NY, USA: ACM, 1999, pp. 93–102.

[25] B. E. John, K. Prevas, D. D. Salvucci, and K. Koedinger, "Predictive human performance modeling made easy," in Proceedings of the SIGCHI conference on Human factors in computing systems, ser. CHI '04. New York, NY, USA: ACM, 2004, pp. 455–462.

[26] A. Cypher, "Eager: programming repetitive tasks by example," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ser. CHI '91. New York, NY, USA: ACM, 1991, pp. 33–39.

[27] A. D'Ulizia, F. Ferri, and P. Grifoni, "A survey of grammatical inference methods for natural language learning," Artif. Intell. Rev., vol. 36, no. 1, Jun. 2011, pp. 1–27.

[28] Human Interfaces in Information Systems (HIIS) Laboratory. ConcurTaskTrees Environment. [Online]. Available: http://giove.cnuce.cnr.it/ctte.html [retrieved: 11, 2014] (2014)

Schema Quality Improving Tasks in the Schema Integration Process

Peter Bellström
Information Systems
Karlstad University
Karlstad, Sweden
e-mail: peter.bellstrom@kau.se

Christian Kop
Institute for Applied Informatics
Alpen-Adria-Universität Klagenfurt
Klagenfurt, Austria
e-mail: chris@ifit.uni-klu.ac.at

***Abstract* — In this article, we address quality in the schema integration process. More specifically, we focus on schema quality improving tasks in the schema integration process. In doing so we describe best practices found in literature used for conceptual modeling as such and apply these to schema integration tasks. Particularly, we address five tasks within the integration process that if used with best quality practices should improve the quality of the integrated schema. Within each best practice, we also address the usage of knowledge repositories to aid in the process of creating a high quality integrated schema. The five tasks are as follows: choosing the right integration strategy, choosing the right conflict resolution methods for the chosen level of abstraction, introducing inter-schema properties to improve and clarify dependencies, combining methods, approaches and guidelines to facilitate recognition of conflicts and finally restructuring. The main contribution is given in the terms of which best practices in conceptual modeling are combined with important integration tasks.**

Keywords - Information and Model Management, Organizational Information, Schema Integration, Schema Integration Process, Schema Quality, Best Practice for Conceptual Modeling

I. INTRODUCTION

Schema integration has a long research tradition. Nevertheless, it is still relevant and many tasks of the schema integration process are needed all the time. There are several reasons and application scenarios, which emphasize this prediction.

Schemata are not built from scratch nowadays. There are many schemata available on the Web. Several of these schemata can be (re)used. Reusing one means that it must be aligned and if needed finally integrated with existing schemata.

Enterprises involve a great deal of data, which constitute an important economic resource and have to be maintained carefully. From an economic point of view, data can be classified into master data, inventory data and transaction data. Especially master data can be used in different information systems within an enterprise and thus shared across this organisation. Integration of data schemata (e.g., schemata of master data) becomes necessary if, for instance, two enterprises merge. A data schema that is used in several enterprises has to be at least aligned too. For another application scenario where integration can take place, we can consider an international operative enterprise with branches in different countries. It has to be expected that the branches will show a tendency to generate proprietary schema parts, important only for this branch. Therefore, the schema will evolve over time. Since the master data schema has to be shared by the whole enterprise, it is necessary to integrate new information consistently into the existing enterprise master data schema.

Finally, if enterprises use provided Web Services, then it might be good to know the business process model and at least match the business process models and data models and check for compliance of the models to the Web Service with the respective models of the enterprise.

A good quality of the result in such contexts is very important. Literature on quality mainly focuses on the quality of the product (i.e., the model). The criteria a model must fulfill in order to have a certain quality is explained. In order to achieve this quality one must also look at the process and think about improving the process tasks.

In [1], we gave a first description of what can be done in the integration process of static schemata in order to get a good integrated model (schema). In this article, we present and describe a continuation of this work. Since an integrated schema is a schema too, we analyzed the literature focusing on static modeling and the kind of process tasks that lead to a better quality of the schema. Then we applied strategies to tasks that have to be done in the integration process. Particularly, we address five tasks within the integration process that if used with best quality practices should improve the quality of the integrated schema. The five tasks are as follows: choosing the right integration strategy, choosing the right conflict resolution methods for the chosen level of abstraction, introducing inter-schema properties to improve and clarify dependencies, combining methods, approaches and guidelines to facilitate recognition of conflicts and finally restructuring.

Since the paper covers the schema integration process and the quality of the integrated schema and the process, this paper is structured as follows. In Section II, we give an overview on integration approaches and quality of schemata. In Section III, we describe the integration process. Section IV focuses on some best practices for improving schema quality. In Section V, we describe the impact of the best practices mentioned in Section IV on the five tasks mentioned in Section III. The paper closes with conclusion and future work.

II. Related Work

In this section, we first address related work in relation to schema integration. This is followed by related work in relation to schema quality. Finally, the section ends with a short summary and a discussion on the research gap that we are addressing.

A. Integration

There is a long research history on several aspects of schema integration. A first substantial work, on integration was made by [2] in the mid-1980s. In another work by [3], other approaches on integration were summarized. In the following years other integration approaches have been published, which focus on several aspects of the integration problem.

In [4], the authors used attribute equivalence as the most basic concept to explain the integration of structural schemata. In [5], the authors presented operators for deciding on the similarity or dissimilarity of schema construct. On the basis of defined assertions, in [6], the author proposed a method to detect equivalent schemata and to automatically integrate two schemata. In [7], the authors concentrate on the automatic detection of naming conflicts. Further algorithms for structural schema integration can be found in [8]. In [9], the authors integrate semantically enriched database schemata. In [10], the author presented an object-oriented framework for the integration of heterogeneous databases. In [11], the authors introduced linguistic knowledge for the integration step. For relationships, for instance, verbs can name relationships. The knowledge about the verbs and their linguistic semantic roles support the integration. In [12], the authors described black board architecture for schema integration of existing databases. With the system, knowledge from designers and end users, which feed the system is shared. The impact of similarity measures for schema matching and data integration is discussed in [13]. In [14], the authors described the integration of state charts in object-oriented models. The work of [15] is based on the formalization of state chart constructs. In [16], the authors proposed a meta-class framework, on which integration should be based. In [17], the author gave an overview of business process integration. In [18], the authors proposed OWL-S ontologies as a support for business process integration. In [19], the authors described the integration of use cases on the basis of petri net models. Finally, in [20] the authors used a behavior tree approach for integrating requirements.

B. Schema Quality

A great deal of work has been done on the quality of conceptual schemata (models) too. Although quality is a feature of a product or artifact (e.g., a schema), it is also necessary to think about the quality of the process of generating the product to support the quality of the product.

In [21], the authors have listed eight schema quality characteristics: completeness, correctness, expressiveness, readability, minimality, self-explanation, extensibility, and normality. In [22], a framework consisting of three dimensions is proposed: "syntax", "semantic" and "pragmatics". The syntax-dimension reflects the vocabulary and grammar (i.e., meta-model) of a schema. The semantic dimension relates the used terms and notions to the domain context. The chosen notions modeled by modeling elements must be legitimate and relevant in the domain, and they must be relevant and legitimate to the purpose for which the schema has been built. Finally, the pragmatic dimension is achieved if the audience can understand and follow the schema.

In [23], the author concluded that there is still a need for standards, which are also accepted by the industry.

In [24], the authors focused on process quality for the development of data schemata (ER diagrams). Their approach was evaluated in a large Australian bank. In the empirical study, it was also important, that the quality was checked throughout the schema development process. In particular, quality-checking was not only made at the end of a phase but before, during and after the schema development phases. Furthermore, it turned out, that an information architect, who checks the schema with respect to enterprise terms, can support the quality.

In [25], the authors presented a framework of four quality characteristics for an ER modeling language: clarity, simplicity, expressiveness, and minimality.

In [26], the authors described the "Guidelines of Modeling (GoM)". Six principles of modeling are introduced in this framework: correctness, relevance, economic efficiency, clarity, comparability and systematic design. These principles can be seen as general strategic and objective definitions for modeling. Based on these goals, the concluded modeling process consisted of the following steps: goal definition, construction of an overall navigation and structural framework, modeling as such, and completion and consolidation.

With the **se**miotic **qual**ity framework (SEQUAL), [27] explains quality of models with model externalization, goals of modeling, modeling domain, explicit knowledge of social actors, interpretation of the social actors and technical actors as well as with language extension.

C. Summary of the Literature

We adopted the integration process as described in [3], since this is a well-established process. They divided the integration process into four phases: pre-integration, comparison of the schemata, conforming the schemata and merging and restructuring. In Section III, we describe this process in more detail.

In Section IV, we continue the description about schema quality according to some selected best practices out of the list of schema quality approaches. We have chosen these approaches since they have shown in practice that they improve schema quality. In Section V, we will then take specific best practices and combine them into five tasks of the integration process steps described in Section III.

III. Integration Process

This section should be viewed as a reference point for the following sections, in which we describe and discuss best practices in the schema integration process. The integration

process starts with a set of schemata, often referred to as views. These views are integrated in order to evolve the global schema. The schema evolvement takes place in the four phases proposed in [3], starting with pre-integration (A), followed by comparison of the schemata (B), and conforming the schemata (C), ending with merging and restructuring (D). The output of one phase is used as the input of the next phase (see Figure 1).

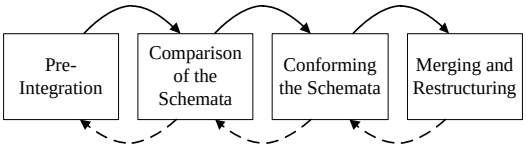


Figure 1. The Schema Integration Process (adapted and modified from [28] p. 20)

A. Pre-Integration

Several tasks should be carried out in the first phase of schema integration. In [29], the author mentions that translating all schemata into the chosen modeling language, checking for differences and similarities in each schema and selecting the integration strategy are all tasks to be performed in pre-integration. Three additional tasks to perform in pre-integration are proposed in [30] as follows: schema element name adoption, schema element disambiguation and introduction of missing relationships.

The output from this phase is a set of revised schemata, the definitions of schema elements and the chosen integration strategy.

B. Comparison of the Schemata

The second phase of schema integration has been researched a great deal and it has been called an important [29] and difficult phase [31][32]. Several authors [3][29][33] have assigned the following tasks to this phase: recognition of name conflicts, recognition of structural conflicts and recognition of inter-schema properties.

The output from this phase is a description of schema element similarities and a description of differences and a description of inter-schema properties.

C. Conforming the Schemata

Also conforming the schemata, phase three has received some attention by other researches. For instance, in [32], the authors called it the most critical phase and in [13] it was called the key issue in schema integration.

In conforming the schemata, the recognized similarities and differences are resolved by adjusting the input schemata.

The recognized inter-schema properties are also used in this phase. However, its full value is shown in merging and restructuring where they are used as a guideline while not only merging the schemata but also restructuring the integrated schema.

The output of this phase is a set of revised schemata.

D. Merging and Restructuring

The last task in the schema integration is merging and restructuring, in which the first task is to merge the revised input schemata into one global intermediate schema. The intermediate schema is then restructured, e.g., detected inter-schema properties are introduced to semantically enrich the schema. Furthermore, schema elements that are truly redundant are recognized and removed from the schema. Merging the schemata as well as restructuring the schema results in a new intermediate schema.

Before the integrated schema is handed over to the developers implementing the information system, the schema is again analyzed, meaning that the schema is checked and verified according to several quality criteria [3][21] and/or quality factors [24].

The result of this phase should be a high quality schema that can be passed to the following phases, in which the information system is implemented.

IV. Some Best Practices Regarding Schema Quality

Both the Guidelines of Modeling (GoM) [26] and the quality factors explained in [24] have a focus on improving the quality of the modeling process and quality of the resulting product (i.e., the conceptual schema).

Both frameworks are a good basis for understanding the quality of the conceptual modeling integration process. The Guidelines of Modeling provide a more strategic framework for covering all aspects of enterprise models (e.g., data, organization, processes, and behavior). The work in [24] focuses on data schemata (models), more specifically schemata modelled with ER diagrams.

Because of its operational focus, we adopted the following practices from [24] for the integration process in order to fulfill the quality factors and improve the quality of the modeling:

- Introducing a specific kind of stakeholder role – the information architect
- Introducing continuous quality checks and reviews.

As well as the general practices:

- Stakeholder participation
- Introducing naming conventions and standards

The information architect (in [24] called data administrator) is a person introduced to review a schema with respect to the other data schemata (models) existing in the enterprise.

According to the authors of [24], who proposed continuous checks and reviews for schema development, reviews must not only be made at the end but also before and during a development step. Such reviews should support the aim of the total quality management that the quality checks and reviews should NOT detect errors but prevent errors.

The participation of different kind of stakeholders is a successful technique used in Information Systems and Enterprise Engineering. Since the schemata (models) represent the knowledge of ideas of people with different backgrounds, it is necessary that different stakeholders are involved.

The introduction of an information architect implies also the usage and management of standards (e.g., what a schema

should look like syntactically, what terms are used and preferred to other terms, etc.).

Beside this, the author of [27] also motivates the language as a factor for schema quality. According to [27] an important means to achieve good schema quality is to choose an appropriate language. He puts the language into the right perspective. A good language is useful but not sufficient. Someone can still generate a poor schema with a good language. Furthermore, the language is chosen already when modeling the original task. But a language has two aspects, one regarding its notions defined in the meta-model and the other aspect is the external representation of notions. In [49], the author proposed nine principles of language representation: Semantic clarity, Perceptual discriminability, Semantic Transparency, Complexity Management, Cognitive Integration, Visual Expressiveness, Dual Coding, Graphic Economy, Cognitive Fit. Semantic clarity means that there must be a one to one mapping between a representation and a notion. Perceptual discriminability is given if concepts are well distinguishable with their representation. A semantic transparency exists if the representation supports the meaning of the notion. With complexity management, the level of abstraction and filtering is supported by notion representations. A language has a cognitive integration if it is possible to navigate between subsets (i.e., different diagrams) of the language. Visual expressiveness describes to what extent the language cognitive variables (e.g., shape, size, color, brightness, etc.) support good interpretation and understanding of a schema. Does the language support graphical notation together with text (i.e., dual coding)? Are there not too many symbols for expressing notions of the language (graphic economy)? Finally, is it possible to adapt and use symbols, which were selected for the specific audience and the skills of the audience (i.e., cognitive fit)? Whereas the language itself cannot be changed, the external representation of the language can be changed in order to fit with the skills of the audience. The minimal change is the harmonization of external representations (semantic clarity), if there is no one-to-one mapping between the external representation and the notion and one notion has more representations. Another possibility would be to transform more abstract representations of notions to representations that fit with the skills of all users. Finally, representations can be changed in size and color to represent a certain state of a schema element (e.g., a concrete class). This would be a way to express that a certain schema element is already integrated.

The use of boundary objects can be another best practice especially for the communication between different kinds of stakeholders. In [34], the authors introduced boundary objects to communicate between professionals and amateurs in the zoological research field. Boundaries are abstracts that support the sharing of the knowledge and communication between communities of practice. Boundary objects are interfaces between these communities. In [35], the author described four classes of boundary objects: repositories, standardized forms and methods, physical objects like prototypes or models. Furthermore, he distinguishes between three types of knowledge boundaries between communities of practices: syntactic, semantic and pragmatic boundaries. Syntactic boundaries exist if different communities have different vocabularies. In this case, a common lexicon can support the overcoming of the differences. To solve semantic boundaries, the parties must create and define a common meaning with the help of boundary objects. Finally, boundary objects for pragmatic boundaries help to establish a negotiation process to find common interests. A pragmatic boundary always includes a semantic and a syntactic boundary. Also, a semantic boundary includes a syntactic boundary. In [36], enterprise architecture artifacts are introduced as boundary objects to support the communication and coordination in an enterprise transformation process.

Boundary objects should also be used during integration since many different kind of stakeholders are involved. Stakeholders can have different views on the domain and even different vocabularies and meanings. They also use parts of the schemata differently. Hence, schema integration has to solve even pragmatic boundaries. Therefore, boundary objects for pragmatic boundaries are needed.

Even the schema in the original modeling language can be a boundary object, because boundary objects can be schemata too. In the following, we will differentiate between a schema, the modeling language already in use for modeling and integration and boundary objects in a stricter sense. In this sense, a boundary object is any extension to the given schema or any additional method or schema from a different modeling language, if the original modeling language does not have sufficient power to act as an interlingua between all stake holders.

However, no predefined set of boundary objects exists, which fulfill the criteria to support communication. Therefore, the challenge is to find the adequate boundary object for a communication purpose. If there were several previous integration process projects in an enterprise, then the stakeholders can rely on given experiences. However, if it is the first integration project, then the stakeholders must agree on what kinds of models, objects, repositories or the like that they will use.

V. Applying Best Practices to Integration Tasks

In general, the best practice of "continuous improvement" is a driver for the whole integration process. Whereas quality is usually considered in or even after the last step of schema integration, we will follow the principle to introduce quality as early as possible. Therefore, we will focus on tasks needed during the whole integration process and not only in the last phases. We will relate the tasks to the best practices in order to improve them. These tasks are: choosing the right integration strategy, choosing the right conflict resolution methods for the chosen level of abstraction, introducing inter-schema properties to improve and clarify dependencies, combining methods, approaches and guidelines to facilitate recognition of conflicts and finally restructuring.

A. Choosing the Right Integration Strategy

In [3], several strategies are proposed to integrate end-user schemata (views). They distinguish between binary and n-ary integration strategies. Among the binary strategies, a ladder strategy or a balance strategy can be chosen. In the ladder strategy, the stakeholders start with two views. They integrate these two views. Then the first integrated schema is compared and matched with another view and so on. In the balanced strategy two views are integrated to become an intermediate schema. This intermediate schema is integrated with other intermediate schemata until the global schema is reached. The n-ary strategies are the one-shot strategy (a global schema is generated at once from all views) and the iterative strategy. The iterative strategy uses one shot strategies only to produce intermediate schemata. These schemata are then integrated with each other (two or more). Integrated schemata can also be integrated with further views. The iterative strategy can be seen as a mixture of the previous three strategies.

Figure 2 illustrates how the binary ladder strategy aids in the process of providing enough points of inspection during the schema integration process.

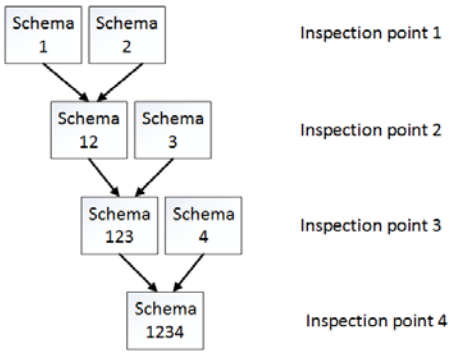


Figure 2. Binary ladder integration strategy with enough inspection points

1) Continuous Checks and Reviews

For continuous checks and reviews, the integration strategy must provide enough definite points of inspections.

A one shot strategy can be excluded as a good strategy by applying this best practice. Otherwise, it would mean that a global schema exists without any intermediate results. If intermediate results are missing, then it is impossible to define definite review milestones. Following the best practice of continuous improvement given in literature, an iterative and balanced or ladder strategy should be applied. Doing so, this intermediate schema can be reviewed each time an intermediate schema is generated.

It cannot be decided which of the other three strategies that should be chosen since all these strategies have intermediate points where schemata can be reviewed before or during integration. The decision between a balanced, a ladder, or an iterative strategy, is a pragmatic decision of available time for the integration and other environmental factors.

2) Information architect, stakeholder participation and standards

Since integration is part of modeling an information architect, standards and stakeholder involvement are also necessary for integration.

The information architect has to assure that a certain intermediate schema as well as the additional views already have to be integrated in compliance with existing schemata in the enterprise. Stakeholders check the semantic correctness and completeness with respect to a certain examined section represented by the views (schemata) or intermediate schemata. For both information architect and stakeholder involvement, strategies that have more intermediate points for discussions and reviews (i.e., ladder, balanced, iterative strategy) are more supportive.

Standards help to check if the schema is syntactically correct and if terms are used in compliance with the enterprise. It is therefore necessary that standards are used. Standards equally drive all the four strategies (one shot, ladder, balanced and iterative). Knowledge repositories, such as stemmers and lemmatizers, could be used to facilitate the task of checking that terms are used in a correct way. Drawing tools might also aid in the modelling process and be used to check that the schema is syntactically correct.

B. Choosing the Right Conflict Resolution Methods for the Chosen Level of Abstraction

In the second phase of schema integration, comparison of the schemata, two schemata are always compared to find the similarities and differences often referred to as conflicts. In the third phase that follows, conforming the schemata, the similarities and differences, i.e., the conflicts, are resolved. The problem is that the same resolution methods are often proposed (and used) for both implementation neutral schemata and implementation dependent schemata. However, using different conflict resolution methods for different levels of abstraction is very important since schemata are designed to be applied on different levels of abstraction. For instance, an implementation neutral schema is often used in the earlier phases of information systems development while an implementation dependent schema in the later phases is close to programming and technical issues.

The purpose of the schema under design may also vary. This is also addressed in [37], in which the authors state that "A schema can serve at least four different purposes. First, it can be used for clarifying the language used in an organisation. Secondly, it can be used for making explicit the rules that prevail in an organisation, which helps to criticise them and possibly to draw up new rules. Thirdly, a schema can be useful for reviewing existing information systems. Fourthly, a schema can be used for developing a new information system." (p. 122).

One way to combine the two levels of abstraction, implementation neutral and implementation dependent, with the quotation given in[37] is described in Table I.

Summing up, if the differences between an implementation independent schema and an implementation dependent schema are ignored and the same conflict resolution methods are used, we might end up with not only a schema that is hard to understand but also end up with semantic loss. It is therefore of great importance that

choosing the right conflict resolution methods for the chosen level of abstraction is taken into consideration in schema integration.

TABLE I. SCHEMA PURPOSE COMBINED WITH SCHEMA LEVELS OF ABSTRACTION

Purpose ([37])	Level(s) of abstraction	Comment
Clarifying the language used in an organisation.	Implementation neutral level	Often the designers are interested in concepts and connections between concepts and not in implementation dependent issues.
Making explicit the rules that prevail in an organisation.	Implementation neutral level	Rules must be expressed in a way that makes it possible for stakeholders to understand the rules and thus be able to criticize them.
Reviewing an already existing information system.	Implementation dependent level	In this case the schema should describe an already implemented information system.
Developing a new information system.	Implementation independent level / implementation dependent level	The designers might not only use different schemata during the development of the information system but also different modeling languages dependent on phase and focus in the information systems development process.

1) Continuous Checks and Reviews

Having compared two source schemata and recognized the conflicts between these, it is important that during the third phase, conforming the schemata, the right conflict resolution methods are used. Since this is not always the case, applying the best practice of continuous checks and reviews are of importance. If the designers and/or stakeholders introduce a resolution method that should not be used for the current level of abstraction, it should be recognized during continuous checks and reviews and changed to the right one. This should in the end contribute to an integrated schema with high quality since an additional check and review has been conducted. For instance, if during comparison of the schemata we have recognized a synonym conflict, e.g., *Customer* in schema 1 and *Client* in schema 2 (see Figure 3), it should be resolved during the conforming of the schemata by introducing a resolution method that is applicable for the current level of abstraction.

It should be noted that if the schemata are designed on an implementation neutral level it is important that a resolution method that retains all concept names and dependencies are applied since they might be of importance for one or several stakeholders. We should therefore not rename one or both concept names, which is one of the most common proposed resolution methods for a synonym conflict but instead introduce a resolution method that keeps both concept names. One way to fulfill this could be to introduce mutual inheritance dependency described as: A and B are synonyms if and only if A inherits B and B inherits A [39].

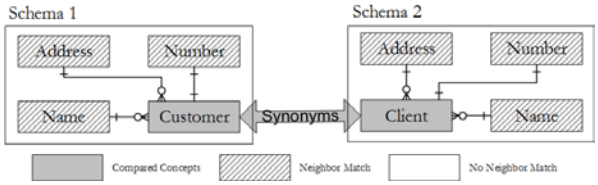


Figure 3. Recognition of synonyms [38] (p. 71)

2) Information architect, stakeholder participation and standards

By involving not only the information architect but also the stakeholders, several of the mentioned pitfalls can be recognized and addressed in the current iteration cycle. This is motivated since it is the stakeholder and the information architect that possess the knowledge about their organization and on how the concepts should be named and connected to each other. The information architect also has to take into account already existing source schemata within the enterprise while a stakeholder might instead focus on integrating a schema of a specific department.

Naming conventions, standards and ontologies, so-called knowledge repositories, might also exist in the enterprise that need to be taken into account in the integration. It should be noted that these naming conventions and standards do not restrict the naming of the concepts which impoverish the language used in the schema but instead are used as a tool to facilitate the integration process as such. Therefore, standards should *not* enforce the usage of one concept name but instead give guidelines on how concept names should be used such as name concepts in the singular.

3) Modeling languages, its external representations and boundary objects

Conventions are not only necessary for the naming of the schema elements. If a language does not have a one to one mapping but a symbol redundancy exits [20], then one and the same symbol has to be chosen in all schemata. Otherwise it confuses the stakeholders.

Some authors, e.g., [40], even eschew the distinction between classes and attributes if possible. The modeling language ORM [40] focuses on the representation of facts. There are no classes and attributes. Instead, object types are related to each other via roles. KCPM [41] has adopted this strategy. This not only helps to be more stable if requirement changes occur, but also has an advantage in schema integration. Problems of structural conflicts can be avoided.

In KCPM, even glossaries were added as an additional means for representing the schema. Such modeling languages and more sophisticated representations of schema elements can be used for both implementation independent and implementation dependent schemata. If the language itself does not provide a glossary representation, it can be introduced as a boundary object. Additionally, ontologies

and knowledge repositories can be seen as boundary objects too.

C. Introducing Inter-Schema Properties to Improve and Clarify Dependencies

Another task in pre-integration (phase 1) and comparison of the schema (phase 2) is the recognition of inter-schema properties. An inter-schema property is not really a conflict but it describes a specific dependency (link) between two concepts often referred to as two concepts that are similar but not exactly the same concept. Two of the most common inter-schema properties as described in literature are holonym-meronym dependencies "part-of" (see Figure 4a) and hypernym-hyponym dependencies "is-a" (see Figure 4b).

Figure 4. Inter-schema holonym-meronym property (a) and inter-schema hypernym-hyponym property (b)

When two schema elements partly match and have been recognized as an inter-schema property, it is documented and passed to the following phase in the schema integration process, in which it is used as a knowledge repository and/or guideline on how to resolve the partly recognized match.

Introducing and being able to use inter-schema properties in the schema integration process is of great importance since an inter-schema property should have a clearer meaning then, for instance, an association dependency between two concepts. The inter-schema property should therefore also be used not only to clarify and improve a specific meaning of two concepts but also to reduce the number of concepts in the integrated schema if possible. Nevertheless, it should be noted that reducing the number of concepts should be done very carefully. Deleting a concept might not only reduce the quality of the integrated schema but also at the worst violate the completeness quality factor addressed in [24].

Finally, a holonym-meronym dependency might be of two types: aggregation and composition, in which composition is the stronger one.

1) Continuous Checks and Reviews

In the second phase of the schema integration process, comparison of the schemata, either the binary strategy (see Figure 2) or n-ary iterative strategy should be used while recognizing similarities and differences, e.g., inter-schema properties, between two source schemata. When an inter-schema property has been recognized, it should be documented and passed on to the following phases in the integration process. In the end, the inter-schema property should not only, be treated as a source to semantic improvement but also be used as guidance anda knowledge repository.

Nevertheless, an inter-schema property should be used in the right way and not in a way that pollutes the source schemata and/or the integrated schema. In the worst case, an inter-schema property is used in a wrong way causing semantic errors. Applying the best practices of continuous checks and reviews is therefore of great importance to improve not only the quality of the integrated schema as such but also to verify that the inter-schema property is used in a correct way.

For instance, if we in the comparison of the schemata have recognized not only a hypernym-hyponym dependency between concept *Article* and *Product* in schema 1 but also a hypernym-hyponym dependency between concept *Product* and *Article* in schema 2, problems might later on be introduced into the integrated schema (see Figure 5).

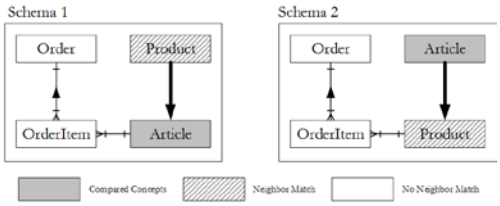


Figure 5. Recognition of difference between two source schemata including inter-schema hypernym-hyponym properties [38] (p. 90)

The inter-schema dependencies are documented and passed on to the following phase, conforming the schemata, in which the schemata are adjusted to solve the recognized conflicts and inter-schema properties. Finally, the modified source schemata (and some extra information resources) are passed to the last phase, merging and restructuring, in which the schemata are first merged and later on restructured. In the worst case, both hypernym-hyponym dependencies described above, *Article* and *Product*, are introduced to the integrated schema causing what is sometimes called reverse subset relationship [21] or cyclic generalization [29] (see Figure 6).

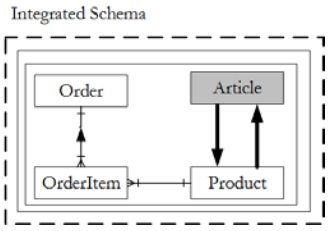


Figure 6. Reverse subset relationship / cyclic generalization (adapted and modified from [38])

However, applying the best practice of continuous checks and reviews, this problem should be recognized and resolved in the current iteration cycle and not left till later iterations in the integration process. Figure 7 illustrates how the reverse subset relationship / cyclic generalization can be resolved by introducing mutual inheritance dependency [39].

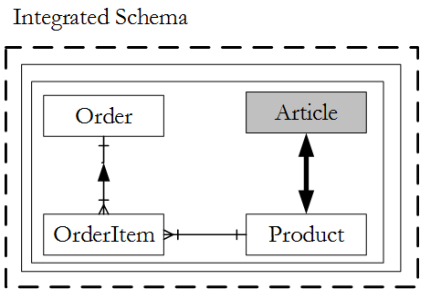


Figure 7. Reverse subset relationship / cyclic generalization resolved by introducing mutual inheritance depencency [38] (p. 90)

2) Information architect, stakeholder participation and standards

Introducing inter-schema properties should result in a semantically richer schema since the inter-schema properties should have a much clearer meaning compared with the association dependency with or without specified cardinality, for instance. Nevertheless, introducing new schema constituents could result in new problems and errors not only since it is the stakeholders that have to be trained in using the new constituent in a correct way, for instance, during a modeling sessions, but also since the new constituent needs to be taken into account during schema integration. Involving information architect as well as stakeholders is also of great importance, since these actors possess the knowledge about their specific domain. Therefore, they also know how to name concepts and how concepts should be connected.

Finally, so-called knowledge repositories (e.g., naming conventions, standards and ontologies) might also exist within the enterprise where the integration is taking place. Ontology, or even domain ontology, might be useful when deciding how to resolve the cyclic generalization dependency. This is the case since a description on how concept *Article* and concept *Product* are dependent might be stated in the ontology (see Figures 5-7).

3) Modeling language and boundary objects

If the modeling language does not provide the possibility to model inter-schema properties, then these dependencies can be seen as a boundary object.

D. Combining Methods, Approaches and Guidelines to Facilitate Recognition of Conflicts

In the first phase, pre-integration, as well as in the second phase, comparison of the schemata, the source schemata are analyzed aiming to recognize similarities and differences within one source schema and between two source schemata, generally referred to as conflicts. In doing so, several matching approaches are needed. The result from each matching approach also needs to be combined into one result. This is motivated since combining the result from several matching approaches into one result should produce a better result than just using the result from one single approach [42]. For instance, in [43], the author has described and exemplified the use of matching approaches for recognizing similarities and differences while integrating structural Karlstad Enterprise Modeling schemata. In [43], the author uses a composite schema based matching approach, in which "[…] the match result of a first matcher is consumed and extended by a second matcher […]" [42] (p. 343) The composite schema based matching approach described in [43] is divided into two parts stating with element level matching followed by structural level matching (Figure 8). Element level matching includes the usage of *concept name comparison*, *linguistic rules* and if a domain ontology exists also *domain ontology-based matching*. Structural level matching includes the usage of *rule-based comparison* and if they exist also *domain ontology matching* and/or *taxonomy based matching*.

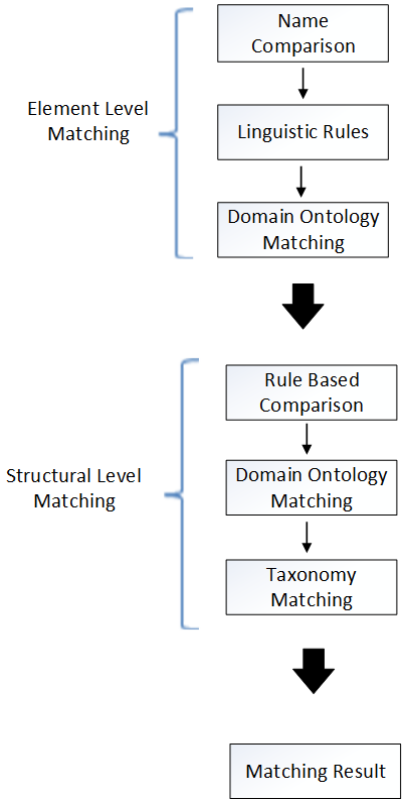


Figure 8. Matching as described and discussed in [43]

To illustrate the importance of using several methods, approaches and guidelines, so-called matchers, we will shortly address some aspects of the matching approach illustrated in Figure 8. In doing so, we use the example schemata illustrated in Figure 9. For a more complete description and discussion of the example, the reader should refer to [43]. It should be noted that in the matching example described below we focus on recognizing similarities and differences between two source schemata. In other words, we emphasise the second phase, comparison of the schemata, mentioned as a challenge [33], also referred to as an important [29] and difficult [31] [32] [44] phase of schema integration. It should also be noted that the end result of the matching approach might include redundant dependencies and concepts since we have not yet decided on what

dependency to use, e.g., synonym or inter-schema hypernym-hyponym.

Figure 9a illustrates the result after conducting the first phase, pre-integration and Figure 9b the result after applying the composite schema based matching approach as described in [43] but before deciding if it is a conflict, synonym, or an inter-schema hypernym-hyponym dependency between Order and High Priority Order. Finally, Figure 9c shows the legend of the used symbols in Figure 9.

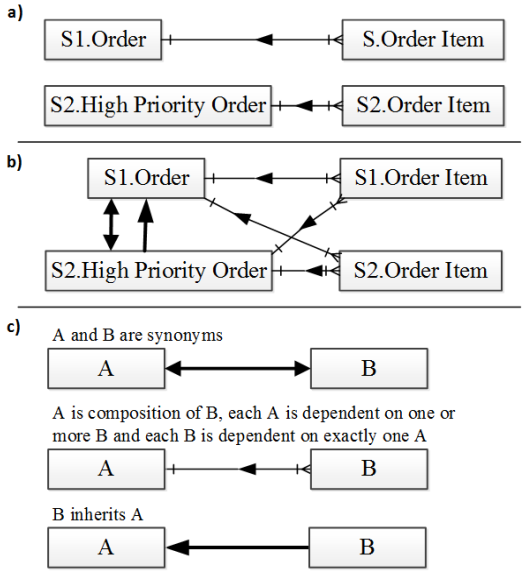


Figure 9. Illustrating example of matching approaches (adapted and modifed from [43])

Figure 9a should be interpreted as follows: *S1.Order* is composed of one or several *S1.Order Item* while *S1.Order Item* is part-of (dependent on) exactly one *S1.Order*. The same interpretation is applicable for *S2* with the change of concept name from *S1.Order* to *S2.High Priority Order*.

Figure 9b should be interpreted as follows: *S1.Order* is composed of one or several *S1.Order Item* and *S2.Order Item* while *S1.Order Item* and *S2.Order item* is part-of exactly one *S1.Order*. *S2.High Priority Order* is composed of one or several *S1.Order Item* and *S2.Order Item* while *S1.Order Item* and *S2.Order item* is part-of exactly one *S2.High Priority Order*. Finally, *S2.High Priority Order* is-a *S1.Order* and *S1.Order* and *S2.High Priority Order* are synonyms.

Focusing on name comparison, linguistic rules and rule based comparison, we may describe the process of matching the two source schemata in Figure 9a as follows:

Name comparison, the labels of schema 1 are compared to the labels of schema 2, on the element level results in the following correspondences: *S1.Order Item = S2.Order Item*, *S1.Order Item ~ S2.High Priority Order*, *S1.Order ~ S2.Order Item* and *S1.Order ~ S2.High Priority Order*.

Applying two linguistic rules on the element level sharpen the meaning of the three last correspondences as follows: *S1.Order Item belongs/related to S2.High Priority Order*, *S2.Order Item belongs/related to S1.Order* and *S2.High Priority Order is-a S1.Order*. The linguistic rules are in [30] (p. 415) described as follows:

*a) If the compared schema elements have names in the form of A and AB […], then the relationship **"AB belongs/related to A"** can be assumed between the elements.*

*b) If the compared schema elements have names in the form B and AB […], then the relationship **"AB is a B"** can be assumed between the elements.*

Applying rule based comparison, first addressed in [38] and later adapted and modified in [30] [43] [45] [46], on the structural level, the following might be suggested: *S1.Order is synonymic with S2.High Priority Order* (1), *S2.High Priority Order is-a S1.Order* (2), *S2.Order Item is part-of (composition) S1.Order* (3), *S1.Order Item is part-of (composition) S2.High Priority Order* (3) and finally *S2.High Priority Order is-a S1.Order* (4).

The rules applied in the presented example might be described as follows:

1) If the comparison of concept names, element level matching, yields match and the comparison of concepts neighborhood, structural level matching, yields partial match, with one concept in each source schemata named differently, then synonymic concepts are most likely recognized.
2) If the comparison of concept names, element level matching, yields match and the comparison of concepts neighborhood, structural level matching, yields partial match, with one concept name named with prior addition to the other one, then an inter-schema hypernym-hyponym property is most likely recognized.
3) If the comparison of concept names, element level matching, yields partially match and the comparison of concepts neighborhood, structural level matching, yields partial match or match, with one concept named with a following addition to the other one, then an inter-schema holonym-meronym property is most likely recognized.
4) If the comparison of concept names, element level matching, yields partially match and the comparison of concepts neighborhood, structural level matching, yields partial match or match, with one concept named with a prior addition to the other one, then an inter-schema hypernym-hyponym property is most likely recognized.

1) Continuous Checks and Reviews

Having designed the source schemata, it is important that in both pre-integration and comparison of the schemata a combination of matching methods, approaches and guidelines are used to recognize not only conflicts within one source schema but also conflicts between two source schemata. In doing so, it is possible to check the quality of the schema after each matching approach has been applied and if necessary also to review the schema. It should, however, be noted that the schemata produced using each

matching method, approach and guideline are intermediate versions of the schemata that are finally going to be integrated. For instance, in the example described above, name comparison results in several matches while the linguistic rules that follow sharpen the meaning of these matches ending up with new intermediate versions of the schemata. Doing continuous checks and reviews after each matching method, approach and guideline has been applied, should contribute to a high quality integrated schema. This is motivated since recognizing problems as early as possible should contribute to a review that is not as cumbersome as identifying problems later on in the integration process resulting in big changes.

2) Information Architect, Stakeholder Participation and Standards

As addressed in [42], the selection of matchers, in our case methods, approaches and guidelines can be made both automatically and manually by a user. However, a generic automated solution process, which selects methods, approaches and guidelines to combine, is difficult to accomplish and besides, a manual selection process is easier to implement. A semi-automatic approach, including both automatic and manual tasks, should therefore be chosen. During such semi-automatic approach, the information architect as well as stakeholders should be very much involved. This is also emphasized in [42], in which the authors state that "[…] user interaction is necessary in any case […]" (p. 343), referring to the process of selection of matchers. This is also in line with [24], who state: "Involvement of all stakeholders in the data modelling process was found to be more important than any other single issue in achieving quality improvements." (p. 646). In general, the information architect and the stakeholders should be involved during the whole integration process meaning that they should also be involved while checking and reviewing each source schemata after each method, approach and guideline has been applied.

Finally, standards such as naming conventions are also important to take into consideration during schema matching. This is motivated since ontologies [47] as well as lexicons such as WordNet [48] are useful in the process of recognizing similarities and differences and should therefore be part of schema matching.

E. Restructuring

Restructuring is the last task within the fourth phase (merging and restructuring). If there is a need to semantically enrich the schema, then detected inter-schema properties can be introduced. Usually, the better the steps and tasks before have been executed, the less has to be done for restructuring. However, especially for the implementation dependent level, restructuring is needed to prune and optimize the resulting schema. It is also necessary if schema alignment and merging were done automatically. Once again a semantic and pragmatic understanding of the terms is required.

1) Continuous checks and reviews

Continuous checks and reviews can be applied as a quality improving instrument to avoid that pruning leads to a schema that is not any longer agreed on by all involved parties. After each major restructuring solution the schema should be checked if its semantic content has not been lost.

2) Information architect

Once again the information architect can act as a mediator between the stakeholders. If necessary, it is his obligation to describe the pragmatics and effects of a schema change. He is also a supervisor for executing several schema checking and restructuring methods and as an organizational interface when working with boundary objects.

3) Stakeholder participation

The more stakeholders from different interest group are involved the more perspectives are considered.

4) Modeling languages and boundary objects

Good modeling languages or additional boundary objects can prevent misunderstandings and errors.

In [49], the author has proposed the introduction of icons and pictures to improve the comprehensibility of conceptual schemata. Others [40] [50] have focused on verbalization of conceptual schemata. Verbalization is a technique, where elements of a schema are translated back to its natural language representation. For a class diagram this means that classes and attributes are mainly translated into nouns and noun phrases. Associations between classes are translated into sentences, which contain the translated classes as the sentence subject and objects. They argued that a conceptual schema, which is transformed back to natural languages sentences representing facts of the domain, is more understandable. Especially, non-computer scientists, who are not familiar with the notions and notations of a modeling language, will be supported with this approach.

For implementation dependent schemata additional techniques adopted from model checking and validation can be used for restructuring. Suppose the schema was developed for an information system, particularly for the database used by the information system. Forms (i.e., user interfaces) can help the stakeholders to understand, which schema elements are necessary for which features of the information system.

Visual query languages [51], which allow the navigation through the conceptual schema, are another way of checking the schema. For the OMT modeling language, [52] has proposed the manual checking of the schema against natural language queries. In [53], a proposal was made on how this can be automated. Advantages and technical problems were discussed. With visual or controlled natural language query languages, manual query checking and/or systems that check the schema automatically based on queries, a restructuring and validation cycle can work as illustrated in Figure 10.

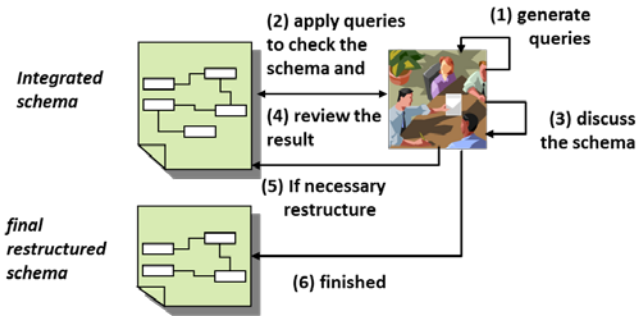


Figure 10. Restructuring task with queries

At the beginning, there is the integrated schema. Since the schema will be implementation dependent, meaning that there will a database working on base of the schema, queries should be generated (1). These are the queries, which most likely will be applied on the integrated database schema. The Information architect together with the stakeholders should generate these queries. Afterwards the queries are executed manually or automatically with a tool on the conceptual schema (2). The result of these executions should be reviewed (2) and the schema should be discussed (3), i.e., is the database schema already optimized for the queries or not. Depending on the result (4), the schema should be restructured and optimized (5) or it turns out that the schema already fits with the retrieval requirements of the stakeholders (6). In this case, the final restructured schema was developed. Finally, with a modeling language (e.g., ORM or KCPM) that does not distinguish between classes and attributes but nevertheless provides mappings to the logical database schema restructuring can be supported. There is no need to prune a class to an attribute if this class does not have further properties. The transformation rules given by these modeling languages will do this.

These examples show that choosing the right modeling language does not only help during the modeling but also later supports possible schema integration. If the original modeling language used is not sufficient, all the stakeholders (i.e., the information architect and the other involved stakeholders) should agree on some boundary objects (e.g., glossaries, or methods such as querying) to get a common understanding and negotiate about the schema.

VI. Conclusion and Future Work

In this article, we have focused on schema quality within the schema integration process. In doing so, we have addressed five best practices of quality improvement given in the literature and five specific integration tasks that should increase the quality of the schema being designed. The best practices addressed are: continuous checks and reviews, information architect, stakeholder participation and standards, a good modeling language as well as the use of boundary objects. The five integration tasks addressed are: choosing the right integration strategy, choosing the right conflict resolution methods for the chosen level of abstraction, introducing inter-schema properties to improve and clarify dependencies, combining methods and approaches and guidelines to conflict resolution and finally restructuring. Wherever it is possible, we have also addressed, within each integration task, how the best practices might be used to aid in the process of producing a high quality schema.

To conclude (see also Table II), the best practices used for conceptual modeling if addressed in connection to schema integration can improve the five mentioned tasks and hence the integration process as such. Continuous checks and reviews, information architect and stakeholder participation can be drivers for choosing the right integration strategy. Continuous checks and reviews, standards, information architect and stakeholder participation are essential in the conflict resolution task. The more conflicts are checked and resolved the better. The more the stakeholders and the information architect is involved the more conflicts can be resolved. Standards support this task as long as they do not restrict the specific naming of enterprise concepts.

For the inter schema property introduction, which is used in at least two phases of the integration process, continuous checks and reviews can help to verify that the inter-schema property is used in the correct way. Stakeholders and the information architect are those who possess the domain knowledge and can thus support the aim to get a semantically richer schema with clear meanings. Standards and ontologies are useful to support the detection of inter-schema properties.

Modeling language of good quality (i.e., language with specific features and a good external representation that supports schema understanding) is important for the following task: choosing the right conflict resolution methods for the chosen level of abstraction, introducing inter-schema properties to improve and clarify dependencies and during restricting. Boundary objects can also be applied in choosing the right conflict resolution methods for the chosen level of abstraction, introducing inter-schema properties to improve and clarify dependencies and during restricting.

In the long run, these improved tasks contribute to a high quality integrated schema.

To summarize, in this work we have described a framework for the schema integration process and aligned best practices to task. In future, we will study some of the best practices in detail. For instance, the kinds of external representations that are good for the schema integration purpose are important to identify. This question is also important for boundary objects. Are there specific boundary objects that should be preferred? It might also be important to determine if the approaches to restructuring the implementation dependent schema can also be applied to implementation independent schemata.

TABLE II. BEST PRACTICES AND INTEGRATION TASKS

Integration tasks/Best Practice	Choosing the Right Integration Strategy	Choosing the Right Conflict Resolution Methods for the Chosen Level of Abstraction	Introducing Inter-Schema Properties to Improve and Clarify Dependencies	Combining Methods, Approaches and Guidelines to Facilitate Recognition of Conflicts	Restructuring
Continuous Checks and Reviews	Are facilitated by the ladder, balanced and iterative integration strategy.	Are the enablers to verify that the schemata illustrate the chosen level of abstraction during the whole integration process.	Are the enablers to verify that the inter-schema properties are used in a correct way during the whole integration process.	Are the enablers to verify that each method, approach and guideline are used in a correct way and that each of these contributes to a more reliable matching result during the whole integration process.	Are the enablers to verify that the schemata are correct after each greater restructuring.
Information Architect	Checks and verifies, from the perspective of the information architect, that an appropriate integration strategy is choosen.	Checks that the choosen conflict resolution methods are in compliance with existing enterprise schemata.	Checks that the introduced inter-schema properties are in compliance with existing enterprise schemata.	Checks that the result from each method, approach, and guideline complies with existing enterprise schemata.	Moderates the restructruing process. The information architect can give information on which pragmatic effects a certain restructuring decision can have.
Stakeholder Participation	Checks and verifies, from the perspective of the stakeholders, that an appropriate integration strategy is choosen.	Checks that choosen conflict resoluton methods are semantically correct and that the schema is complete from a stakeholder perspective.	Checks that the introduced inter-schema properties are semantically correct and that the schema is complete from a stakeholder perspective.	Checks that each method, approach and guideline produces semantically correct results and that the schema is complete from a stakeholder perspective.	Improves the schema quality since the stakeholders are informed about the effects of a restructing decision and therefore also can influence how restructuring is performed.
Standards	Help in the process of checking that the schemata are syntactically correct and that terms are used in compliance with the enterprise schemata.	Help in the process of introducing the correct resolution method for not only naming conflicts but also structural conflicts.	Help in the process of introducing the correct inter-schema propterty and help in the process of introducing the inter-schema property in a correct way.	Help in the process of introducing and applying each method, approach and guideline.	Not applicable
Modeling language quality	Not applicable	A modeling language, which does not distinguish between classes and attributes can prevent structural conflicts	Not applicable	Not applicable	The modeling language and its external representation can support restructuring (e.g., by providing modeling elements that make restructuring easier)
Boundary objects	Not applicable	With the right boundary objects, the stakeholders can be made aware that conflicts exist.	Interschema properties can be seen as boundary objects.	Not applicable	Boundary objects in the form of additional methods can help to identify possible errors and can support the understanding of the schema.

REFERENCES

[1] P. Bellström and C. Kop, "Towards Quality Driven Schema Integration Process Tasks," Proceedings of the The Sixth International Conference on Information, Process, and Knowledge Management (eKNOW 2014), 2014, pp. 98-104.

[2] C. Batini and M. Lenzerini, "A Methodology for Data Schema Integration in the Entity-Relationship Model," IEEE Transactions on Software Engineering, vol. 10 (6), 1984, pp. 650-664.

[3] C. Batini, M. Lenzerini, and S. B. Navathe, "A Compartive Analysis of Methodologies for Database Schema Integration," ACM Computing Surveys, 1986, vol. 18 (4), pp. 323-364.

[4] J. A. Larson, S. B. Navathe, and R. Elmasri, "A Theory of Attribute Equivalence in Databases with Application to Schema Integration," Transactions on Software Engineering, vol. 15 (4), 1989, pp. 449-463.

[5] A. Savasere, A. Sheth, and S. Gala, "On Applying Classification to Schema Integration," Proceedings of the First International Workshop on Interoperability in

Multidatabase Systems (IMS'91), IEEE Press, 1991, pp. 258-261.

[6] P. Johannesson, "A Logical Basis for Schema Integration," Third International Workshop on Research Issues on Data Engineering (RIDE-IMS'93), IEEE Press, 1993, pp. 86-95.

[7] H. K. Bhargava and R. M. Beyer, "Automated Detection of Naming Conflicts in Schema Integration: Experiments with Quiddities," Proceedings of the 25th Hawaii International Conference on System Sciences, IEEE Press, 1992, pp. 300-310.

[8] J. Geller, A. Mehta, Y. Perl, E. Neuhold, and A. Sheth, "Algorithms for Structural Schema Integration," Proceedings of the Second International Conference on Systems Integration (ICSI'92), IEEE Press, 1992, pp. 604-614.

[9] M. García-Solaco, F. Saltor, and M. Castellanos, "A Structure Based Schema Integration Methodology," Proceedings of the Eleventh International Conference on Data Engineering, IEEE Press, 1995, pp. 505-512.

[10] H. Dai, "An Object-Oriented Approach to Schema Integration and Data Mining in Multiple Databases," Proceedings on the Technology of Object-Oriented Languages (TOOLS), IEEE Press, 1997, pp. 294-303.

[11] E. Métais, Z. Kedad, I. Comyn-Wattiau, and M. Bouzeghoub, "Using Linguistic Knowledge in View Integration: Toward a Third Generation of Tools," Data & Knowledge Engineering, vol. 23 (1), 1997, pp. 59-78.

[12] S. Ram and V. Ramesh, "A Blackboard-Based Cooperative System for Schema Integration," IEEE Expert, 1995, vol. 10 (3), pp. 56-62.

[13] S. Spaccapietra and C. Parent, "View Integration: a Step Forward in Solving Structural Conflicts," IEEE Transactions on Knowledge and Data Engineering, vol. 6 (2), 1994, pp. 258-274.

[14] H. Frank and J. Eder, "Towards an Automatic Integration of Statecharts," International Conference on Conceptual Modeling (ER 1999), 1999, pp. 430-444.

[15] B. H. C. Cheng and E. Y. Wang, "Formalizing and Integrating the Dynamic Model for Object Oriented Modeling," IEEE Transactions on Software Engineering, vol. 28 (8), 2002, pp. 747-762.

[16] M. Stumptner, M. Schrefl, and G. Grossmann, "On the Road to Behavior-Based Integration," Proceedings of the 1st APCCM Conference, 2004, pp. 15-22.

[17] A. Raut, "Enterprise Business Process Integration," Conference on Convergent Technologies for Asia-Pacific Region, IEEE Press, 2003, pp. 1549-1553.

[18] S. Fan, L. Zhang, and Z. Sung, "An Ontology Based Method for Business Process Integration," International Conference on Interoperability for Enterprise Software and Applications in China, IEEE Press, 2008, pp. 135-139.

[19] W. J. Lee, S. D. Cha, and Y. R. Kwon, "Integration and Analysis of Use Cases Using Modular Petri Nets in Requirements Engineering," IEEE Transaction of Software Engineering, vol. 24 (12), 1998, pp. 1115-1130.

[20] K. Winter, I. J. Hayes, and R. Colvin, "Integrating Requirements: The Behavior Tree Philosophy," 8th IEEE International Conference on Conference on Software Engineering and Formal Methods (SEFM), IEEE Press, 2010, pp.41-50.

[21] C. Batini, S. Ceri, and S. B. Navathe, Conceptual Database Design an Entity Relationship Approach. Redwood City: Benjamin/Cummings Publishing Company, 1992.

[22] O. L. Lindland, G. Sindre, and A. Solvberg, "Understanding Quality in Conceptual Modeling," IEEE Software, vol. 11 (2), 1994, pp. 42-49.

[23] D. L. Moody, "Theoretical and Practical Issues in Evaluating the Quality of Conceptual Models: Current State and Future Directions," Data & Knowledge Engineering, vol. 55 (3), 2005, pp. 243-276.

[24] D. L. Moody and G. G. Shanks, "Improving the Quality of Data Models: Empirical Validation of a Quality Management Framework," Information Systems Journal, vol. 28 (2), 2003, pp. 619-650.

[25] S. S. Cherfi, J. Akoka, and I. Comyn-Wattiau, "Perceived vs. Measured Quality of Conceptual Schemas: An Experimental Comparision," Proceedings of Tutorials, Posters, Panels and Industrial Contribution of the Twenty-Sixth International Conference on Conceptual Modeling (ER 2007), vol. 83, 2007, pp. 185-190.

[26] J. Becker, M. Rosemann and C. von Uthman, "Guidelines of Business Process Modeling," Business Process Management, LNCS 1806, 2000, pp. 30-49.

[27] J. Krogstie, Model based Development and Evolution of Information Systems – A Quality Approach. London: Springer, 2012.

[28] P. Bellström, View Integration in Conceptual Database Design Problems, Approaches and Solutions. Licentiate Thesis. Karlstad Univeristy, Karlstad University Press 2006:5, 2006.

[29] W. Song, Schema Integration – Principles, Methods, and Applications. Dissertation. Stockholm: Stockholm University & The Royal Institute of Technology No. 95-019, 1995.

[30] P. Bellström and J. Vöhringer, "A Semi-Automatic Method for Matching Schema Elements in the Integration of Structural Pre-Design Schemata," International Journal on Advances in Intelligent Systems, vol. 4 (3 & 4), 2011, pp. 410-422.

[31] L. Ekenberg and P. Johannesson, "A Formal Basis for Dynamic Schema Integration," Conceptual Modeling – (ER'96), LNCS 1157, 1996, pp. 211-226.

[32] L. Lee and T. W. Ling, "A Methodology for Structural Conflict Resolution in the Integration of Entity-Relationship Schemas," Knowledge and Information Systems, vol. 5 (2), 2003, pp. 225-247.

[33] P. Johannesson, Schema Integration, Schema Translation, and Interoperability in Federated Information Systems. Dissertation. Stockholm University & Royal Institute of Technology No. 93-010-DSV, 1993.

[34] S. L. Star and J. R., Griesemer, "Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkley's Museum of Vertebrate Zoology, ", 1907-39, Social Studies of Science, 19 (3), 1989, pp. 387-420.

[35] P. R. Carlile, "Transferring, Translating and Transforming. An Integrative Framework for Managing Knowledge Across Boundaries," Organization Science, 15 (5), 2004, pp. 555-568.

[36] R. Abraham, "Enterprise Architecture Artifacts As Boundary Objects – A Framework of Properties," Proceedings of the 21st European Conference on Information Systems (ECIS), 2013, Paper 120.

[37] M. Boman, Jr. J. A. Bubenko, P. Johannesson, and B. Wangler, Conceptual Modelling. London: Prentice Hall, 1997.

[38] P. Bellström, Schema Integration How to Integrate Static and Dynamic Database Schemata. Dissertation. Karlstad University, Karlstad University Studies 2010:13, 2010.

[39] R. Gustas, Semantic and Pragmatic Dependencies of Information Systems. Kaunas: Kaunas Technologija, 1997.

[40] T. Halpin and M. Curland, "Automated Verbalization for ORM-2," Proceedings of OTM 2006 Workshop , LNCS Vol 4278, 2006, pp. 1181-1190.

[41] H.C. Mayr and C. Kop, "Conceptual Predeign – Bridging the Gap between Requirements and Conceptual Design,"

Proceedings of the 3rd Int. Conference on Requirements Engineering (ICRE'98), IEEE Press, 1998, pp. 90-100.

[42] E. Rahm and P.A. Bernstein, "A Survey of Approaches to Automatic Schema Matching," The VLDB Journal, vol. 10, 2001, pp. 334-350.

[43] P. Bellström, "A Semi-Automatic Approach for the Integration of Structural Karlstad Enterprise Modeling Schemata," Advances in The Human Side of Service Engineering, 2014, pp. 13-24.

[44] A. Doan, F. N. Noy and A. Y. Halevy, "Introduction to the Special Issue on Semantic Integration," SIGMOD Record, vol 33 (4), 2004, pp. 11-13.

[45] P. Bellström, "A Rule-Based Approach for the Recognition of Similarities and Differences in the Integration of Structural Karlstad Enterprise Modeling Schemata," The Practice of Enterprise Modeling, 2010, pp. 177-189.

[46] P. Bellström and J. Vöhringer, "A Three-Tier Matching Strategy for Predesign Schema Elements," The Third International Conference on Information, Process, and Knowledge Management (eKNOW 2011), 2011, pp. 24-29.

[47] T.R. Gruber, "A Translation Approach to Portable Ontology Specifications," Knowledge Acquisition, 5, 1993, pp. 199-220.

[48] G.A. Miller, "WordNet: A Lexical Database for English," Communication of the ACM, 38 (11), 1995, pp. 39-41.

[49] D. Moody, "The 'Physics' of Notations: Toward a ScientificBasis for Constructing Visual Notations in Software Engineering," IEEE Transaction on Software Engineering, vol. 35 (6), 2009, pp. 756-779.

[50] H. Dalianis, "A Method for Validating a Conceptual Model by Natural Language Discourse Generation," Proceedings of the Fourth International Conference CAiSE'92 on Advanced Information Systems Engineering, LNCS Vol 594, 1992, pp. 425-444.

[51] K. Järvelin, T. Niemi, A. Salminen, "The Visual Query Language CQL for Transitive and Relational Computation," Data & Knowledge Engineering, vol. 35, 2000, pp. 39-51.

[52] J. Rumbaugh, M. Blaha, W. Premelani, F. Eddy and W. Lorensen, Object oriented Modeling and Design. Prentice Hall International Inc. Publ. Comp. 1991.

[53] C. Kop, "Checking Feasible Completeness of Domain Models with Natural Language Queries," Proceedings of the 8th Asia-Pacific Conference on Conceptual Modeling, vol. 130, 2012, pp. 33- 42.

Touch Recognition Technique for Dynamic Touch Pairing System and Tangible Interaction with Real Objects

Using 3D Point Cloud Data to Enable Real Object Tangible Interaction

Unseok Lee and Jiro Tanaka
Department of Computer Science
University of Tsukuba
Tsukuba, Ibaraki, Japan
{leeunseok, jiro}@iplab.cs.tsukuba.ac.jp

***Abstract*— Sensor-based pairing technology between digital objects for interactions is used widely (e.g., smart phone to Bluetooth headset). In addition, research about tangible interactions between daily normal analog objects (e.g., a doll, Lego block) and digital objects has progressed and is also popular. However, such research can only involve interactions with already setup objects. They have to attach sensors to objects for interaction. The paired objects cannot be changed dynamically. In addition, it is difficult to make interactions with various objects simultaneously. The objects with attached sensor(s) for tangible interaction can recognize the touched area of the objects, but cannot recognize touch gesture with a lot of movements. In this paper, we propose a new analog-digital object pairing method and touch recognition technique by intuitive touch interactions using three-dimensional point cloud data. Several touch pairing and touch recognition methods are described in detail. The paired objects are changed dynamically using the proposed method. In addition, tangible interactions between two objects are described after pairing. Finally, we demonstrate the high recognition rate of the proposed method using experiments and describe our system's contribution.**

Keywords-dynamic pairing, point cloud, tangible interaction, 3d gesture, human computer interaction, touch recognition

I. INTRODUCTION

In everyday life, the touching action is natural and common. We touch objects to use them (e.g., a doll or toy to play, open a bottle cap for drinking). Touch interactions with digital devices have also become natural in recent years, because smart devices with touch screens and touch pads are now used widely. Simultaneously, in the field of human–computer interaction (HCI), research on interactions between physical objects and digital devices has progressed rapidly. A physical object is set as an input unit and the digital device is controlled by it. Such interactions are used widely and have become a 'natural' method. However, to use a physical object as an input unit, much effort and time is initially needed to set up sensors [1]. Moreover, it takes time and effort to apply sensors again when using another physical object as the input unit. In addition, the digital object is limited to a particular physical object. The touching objects recognition is also limited. Recognizing the touched area of an object and the touch gesture with 3D objects is difficult by attaching sensors. Thus, there is no 'natural' interaction between various objects. Regarding the input unit, research on methods for making a tangible object for which touch sensing is possible has progressed.

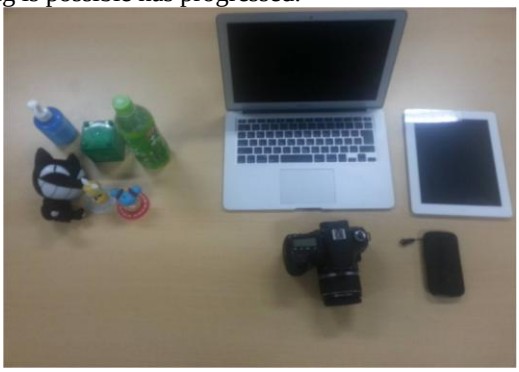


Figure 1. Analog-Digital objects in everyday life

For example, in the bowl project [2], a simple media player in a bowl sits on a living room table and a range of physical objects can be placed within it. When an object is placed in the bowl, related media are played on the TV. The project used radio-frequency identification (RFID) sensors for tangible interactions. However, the system could not provide dynamic pairing and touch gesture recognition between objects. The interactions and possible objects were also limited. The "HandSense" [3] prototype used capacitive sensors for detecting when it was touched or held against a body part. It could determine whether a device was held in the left or right hand by measuring the capacitance on each side. Wimmer [4] presented a method for prototyping grasp-sensitive surfaces using optical fibers. However, all of these examples require attaching sensors to the devices. This is unnatural in the real world. They cannot support multiple object recognition and touch gesture based interactions. Also, the paired object cannot be changed dynamically.

In this paper, we propose a new method for dynamic pairing and touch recognition techniques with tangible interactions between analog objects and digital objects in practical circumstances.

This dynamic pairing is designed through touch interactions. For example, one hand grasps a doll, an analog

object. Then, the other hand grasps a smart phone, a digital object. The doll and smart phone are paired and prepared for tangible interactions. The system makes it possible to pair the doll with touch in three dimensions. The smart phone then shows feedback from interactions with the doll. We can change the paired objects dynamically. Figure 1 shows examples of pairing analog and digital objects in everyday life. We also designed touch recognition technique and learned touch gesture based interaction for natural and robust use with existing objects. The natural actions that can be used as gestures, and supporting all objects in everyday life, were considered for tangible interaction. The touch patterns, practical interactions with popular gestures, object size and object hardness were considered as well.

Our proposed methods and techniques are based on three-dimensional(3D) point cloud data using two Kinect units. They capture and calibrate 3D point cloud data. Our system determines touch pairing and tangible interactions of the paired analog object, based on these calibrated data. In this way, the system can readily recognize what objects are touched and trace what objects are paired. Our system also determines touch recognition and learned gesture based tangible interactions of the object. In addition, we can recognize the touched position and movements of the objects. We present the results of tests of recognition rate for pairing and touch recognition rate using the proposed method.

The rest of this paper is organized as follows. Section II introduces related work on depth-based touch sensing and tangible interactions. We describe in detail the principles of the pairing method and the system specifications in Section III. We described touch recognition techniques in detail and tangible interactions in Section VI. We present details on the high recognition rate of our system in Section V. Finally, we describe our contribution and future work in Section VI.

II. RELATED WORK

A. Depth-based Touch Sensing Technologies

In recent years, depth-based cameras and related technology have developed rapidly. Research on obtaining 3D data on objects using depth information has also progressed. The framework for 3D sensing using depth cameras has been improved remarkably [5].

Klompmaker et al. implemented tangible interactions using a depth camera and a 3D sensing framework [12]. They implemented touch detection and object interaction, supporting multi-touch and tangible interactions with arbitrary objects. They used images from a depth camera to determine whether a user's finger touched the object. However, they were unable to support 3D touching and dynamic pairing between objects for tangible interactions.

Wilson et al. used depth-sensing cameras to detect touch on a tabletop [7], using the camera to compare the current input depth image against a model of the touch surface. The interactive surface need not be instrumented in advance for the interaction and this approach allows touch sensing on non-flat surfaces. However, they only supported simple touch recognition and could not address touch in any direction with 3D objects.

B. Tangible Interactions with Analog Objects

"Digital Desktop" by Wellner et al. [8] was used in an early attempt to merge the physical and digital worlds. They implemented a digital working space on a physical desktop where physical paper served as an electronic document. The interaction with papers was by means of bare fingers. "Icon Sticker" [9], based on this idea, is similar. Icon Sticker is a paper representation of digital content. It consists of transferring icons from the computer screen to paper, so they can be handled in the real world and used to access digital content directly. An icon is first converted into a corresponding barcode, which is printed on a sticker. Then the sticker can be attached to a physical object. To access the icon, the user scans the barcode on the sticker with a barcode scanner. "Web Sticker" [10] uses barcodes to represent online information. It is similar to Icon Sticker, but instead of icons it manages Web bookmarks. They use a handheld device with a barcode-reading function to capture the input and display related information.

There were also attempts to improve tagging of physical objects for a more natural tangible interaction. Nishi et al. [11] registered real objects on a user's desktop based on a user indicating a region on the desk by making a snapshot gesture with four fingers. A color histogram was used to model the object and a pointing gesture was used to trigger the recognition. "Enhance Table" also uses a color histogram to model objects. However, the size is predefined and the system is limited to mobile phone recognition.

Although many previous tangible interaction studies have used physical objects for interactions, most of them are token-based approaches and provide only limited use of real objects. They do not support 3D object tracking or pairing for tangible interactions. Thus, to overcome this, we propose a robust 3D object-tracking method that detects touch in three dimensions. The system supports dynamic pairing between analog and digital objects, and makes analog objects accessible to touch anywhere.

III. TOUCH PAIRING SYSTEM

A. Hardware and Software

Our system consists of two Microsoft Kinect sensors for Xbox 360 with stands.

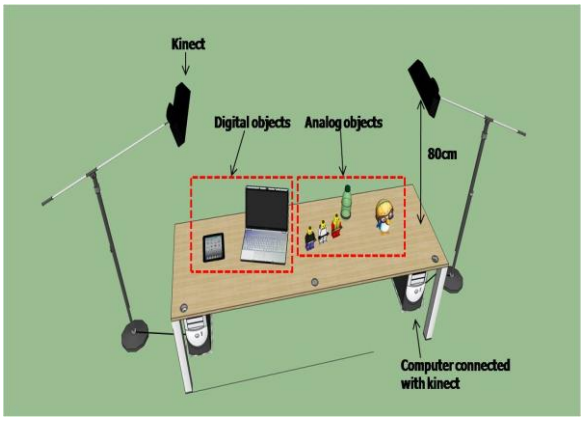


Figure 2. Touch Pairing System Configuration

Two computers (Intel i7 2.4Hz, 8GB RAM, and GeForce GTX 660M graphics card) are used to handle the 3D data. A desk and the pairing object for interaction are installed. The analog and digital objects are randomly placed. The Kinect sensors are set at 80 cm from the desk. The computers are connected to each Kinect sensor, and the digital objects have wireless internet or Bluetooth connections with the computer, installed in the bottom of the desk (Figure 2). Our system uses OpenFrameworks OpenNI [6] for the Kinect sensors and a point cloud method that provides example add-ons of frameworks. The system obtains 3D point cloud data and maps the RGB data to the point cloud. The movement of the points is based on pairing recognition. The proposed system was implemented on a Microsoft Windows 7 platform. The pairing recognition module was implemented in Visual Studio 2010 and OpenNI 1.5.4.

B. *Touch Pairing System Architecture*

The entire system consists of three major modules (Figure 3). The input data are obtained by the two Kinect sensors, on the left and right. It calibrates their data from two cameras and processes the data. In the case of using one camera, there are parts that cannot be reconstructed such as both sides of a cup. Because one camera cannot catch all area of objects even if camera sets up top-down direction. It is difficult to find out the touched location and touching gesture itself. It cannot cover all sides of objects using only one camera.

Our system sets up two cameras in proper position that was found out empirically (Figure 2). The system implemented the calibration and 3D reconstruction to obtained data from two cameras. We can recognize touching almost all sides when using these methods. We can find the location of the touch as well. In addition, the difficult side to recognizing is estimated using reconstructed 3D model data. The process is detailed below.

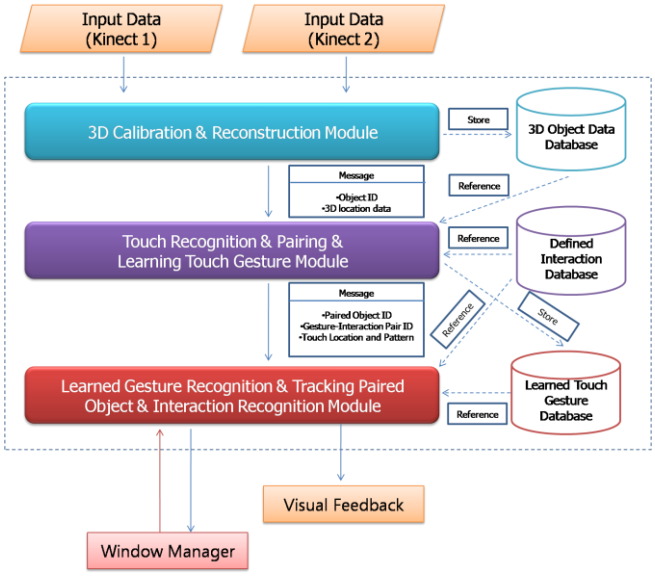


Figure 3. Touch Pairing System Architecture

1) 3D calibration and reconstruction module: In this module, we calibrate depth data for each object, obtained from the two Kinect sensors. The system makes a 3D reconstruction using a point cloud library with calibrated data. The module stores calibrated and reconstructed data in a database, which is then used by the touch-recognition module. After storage, the module sends messages to the touch-recognition and pairing module about object ID and object location using the 3D point cloud data.

2) Touch recognition, pairing, and learning touch gesture module: In this module, we implemented pairing and learning touch gesture with touch recognition method. The process is detailed as follows.

a) Touch recognition and pairing: Touch is recognized in terms of the depth and position of the object and hands using 3D tracking. Using the previous depth information from the 3D reconstruction based on the point cloud, the system determines whether the hand touched the object, and if so, the position of the object. The system recognizes the time of touching between the user's 3D hand point cloud data and the object 3D point cloud data, then determines whether they are paired. A paired analog object's 3D point cloud data are stored and sent to the tracking module with information on the object type. We defined a limited objects database.

b) Learning touch gesture: In this method, the system identificates touched position of hands and fingers with objects. In addtion, our system stores movement of the fingers and gestures to the database. A touched object is stored with its 3D vision image for tracking after paired. Gestures and interaction pair with touched objects can be stored in the database by the system as well. They can be used tangible interactions with learned touch gestures.

3) Learned gesture recognition, tracking paired object, and interaction recognition module: In this module, our proposed system recognizes learned touch gesture and tracks paired object. In addition, a tangible interaction is implemented with touched position and gestures using data from the database. The process is detailed as follows:

a) Learned gesture recognition: The learned touch gesture recognition is implemented using learned touch gesture database that was stored by the learning touch gesture module. The system recognizes touched position and gestures applied to the objects. The paired object and interactions are prepared after recognizing the touch gesture and the touched objects.

b) Tracking paired object and interaction recognition: After pairing, the system tracks the analog object based on saved 3D point cloud data. The paired digital object can be tracked; however, the paired objects do not commonly move. The paired analog object is tangible, based on 3D analog object data, from the 3D calibration and reconstruction module. We can make an interaction with the digital object in this module; the interaction is shown by visual feedback.

C. Touch Pairing Method using 3D Point Cloud Data

Our proposed method uses 3D point cloud processing of Kinect depth data. A point cloud itself is a set of data points in a coordinate system. We measured a large number of points on the surface of an object using OpenFramework [6].

Figure 4. 3D Point Cloud Data

The system obtains RGB data from two Kinect sensors (Figure 4) and assigns them to the depth area. However, not all directions of the object can be reconstructed. Thus, we find the most appropriate location for the Kinects and position them so that they cover most of the experimental space.

1) Touch gesture and recognition: Our touch pair system recognizes the touch actions of users' hands based on depth. The system calculates the depth of each point between a user's hand and the object by filtering closer data. The flow of recognition is as follows.

a) Calculation of all depth points: Calculate all points of the analog and digital objects and the user's hand on the table.

b) Determination of finger position: The system calculates the minimum and maximum depths of the finger by defined thresholds because we hold our fingers in specific ways when we touch something.

c) Determination of hand position: The system calculates the minimum and maximum depths of all fingers and the palm; from the front view, the system uses depth information from both sensors simultaneously.

d) Determination of grasping: Using depth data on users' hands and on objects collected from both sensors simultaneously, we found certain threshold values for recognizing the act of grasping.

2) Analog-digital object pairing: Our system performs time calculations between touched analog-digital objects versus touched object-object pairing. Implementation of object-object touching is shown in Figure 5c,d. When the system recognizes the objects that the user wants to pair, the color is changed. The red color refers to the digital object and the analog object is blue. The recognized hand is shown in yellow using 3D point cloud data. The main steps are as follows.

a) Time calculation: For pairing, the user maintains a touching posture for a few seconds after touching is recognized between the objects. The color is changed after the pairing.

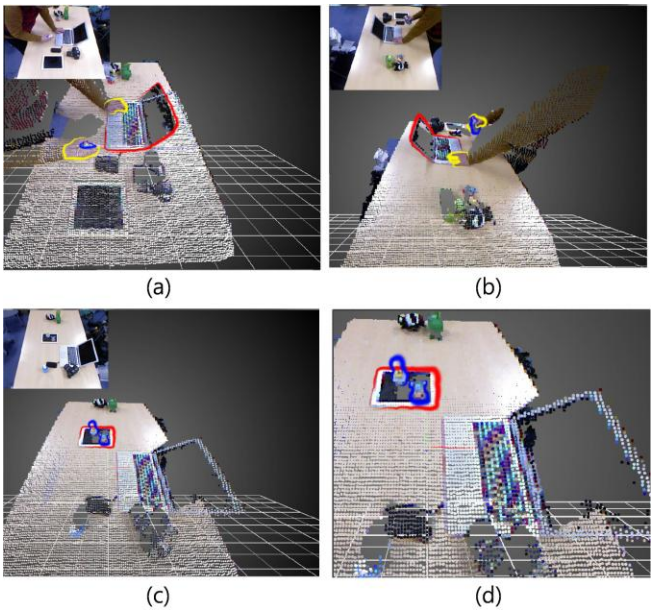


Figure 5. (a) Pairing gesture with hand touching to digital object and grasping analog object from right camera (b) Pairing gesture from left camera (c) Pairing by object-object touching, toy and smart pad are pairing (d) Detailed object-object pairing image

b) Tracking a paired set: To track paired objects, the system calculates 3D point cloud data continuously, which are provided by the real-time reconstruction module (Figure 3). The color of the tracked object is shown.

c) Changing a paired set: To change a paired object set, a pairing gesture is made for some period of time. The user touches what he/she wants to pair. After a few seconds, performing the pairing gesture (Figure 5) will change the pair set as indicated by the color feedback.

IV. Touch Recognition Techniques

In this section, we describe about recognition technology touching 3D object. Our system classified a kind of hand touch patterns. In addition, our system determines whether the object is touched or not in real time. We illustrate about the method recognizing the location of the touched part as well. The system uses RGB-D images from two Kinect cameras and point cloud data for recognizing.

A. Touch patterns of object

The pattern of the touch is various and it is different according to each research. In this paper, we classify patterns into three types of touch in order to use tangible interaction with existing object. Our system classifies into finger touch, hand touch and grasp. Finger and hand touch are top-down touch (Figure 4). Finger touching only or finger and palm touching are distinguished. The above mentioned touch patterns are quite simple touch patterns. On the other hand, grasp touch is more complicated and has various patterns. It

happens when we pick the object up or hold it. For example, picking a pencil and grasping a P.E.T bottle are considered grasp touches. Grasp touches can be classified according to object's hardness and size. We touch fingers and palm to the object when we pick the cup up and touch fingers only when we pick the pencil up. A grasp touch can be classified as finger touch or hand touch.

In this research, the system stores user's touched pattern and location value for making tangible objects from existing objects. The touch patterns are classified as described above and become the fundamental initial point for the interaction.

B. Touch Recognition Flow

Our proposed system classifies three parts (Figure 3). The system performs these methods in order. First is 3D calibration and reconstruction. The system performs calibration process on the data from RGB-D camera1 and camera2. The data from RGB-D camera2 rotates 180 degree based on y-axis because camera1 and camera2's data location are opposite.

The system makes points in each data based on 3D model and vision image when the system performs rotating (see red points and blue points in Figure 6b). An object id is generated using 3D shape and vision image data in real time. The system can recognize 3D location using point cloud data. We can store the touched position of the object by using these data. The stored database consists of a structure similar to Table I. The system reconstructs hands model using point cloud data (see Figure 6e) from two cameras. The hand index is assigned to each hand. The mesh and depth value of each finger in hands are stored to database. It calibrates objects using vision and depth image from two cameras for generating the object id. There is 3D value such as mesh, depth as well from each camera. It provides 3D shape, location and vision image of each objects. The system can discriminate objects using these data. In result, the stored 3D object database is used for touch recognition, touched position of object and learning gesture as well. The second is touch recognition and gesture learning. The touch recognition recognizes by using the point cloud and reconstructed 3D model estimation. Our system uses point cloud data when we do simple touch recognition. It is that case of finger touch or hand touch. The system recognizes the location of hand and objects using depth based point cloud data. We defined the depth value of an object surface as $D_{surface}$. It defined the maximum value as D_{max} that is slightly above the $D_{surface}$. It defined upper part of finger joint as D_{min} as well.

TABLE I. 3D OBJECT DATABASE STRUCTURE

	Kinect 1	**Kinect 2**
Object	Hand Index (Left or Right hand) Hand Value(Mesh, Depth value) Calibrated Object Vision Image and ID	
	Mesh value Depth value Vision image	Mesh value Depth value Vision image

TABLE II. LEARNED TOUCH GESTURE DATABASE STRUCTURE

	Kinect 1	**Kinect 2**
Object	Hand Index (Left or Right hand)	
	Finger [0][0,1,2](Thumb) Finger [1][0,1,2] (Index) Finger [2][0,1,2](Middle) Finger [3][0,1,2](Ring) Finger [4][0,1,2](Little) Palm value Vision image	Finger [0][0,1,2](Thumb) Finger [1][0,1,2] (Index) Finger [2][0,1,2](Middle) Finger [3][0,1,2](Ring) Finger [4][0,1,2](Little) Palm value Vision image

TABLE III. DEFINED INTERACTION DATABASE

	Object
Interaction	Interaction type Gesture-Function Pair ID Touch Pattern Vision image

The system determines when users touch the object surface when the finger joint is between D_{max} and D_{min} [16]. The touched area is between hand and object point cloud data (see red points in Figure 6d). The hand touch is recognized because all finger joints value and palm value are between each $D_{surface}$ value. These recognized and touched data can be stored when the user wants to make touch gestures. The database that is stored by our system consists of structure similar to Table II for learning touch gestures. The table value is stored when the user makes gesture with touching object. The hand index in the Table II means discriminating touched hand with object. The touched finger value with object are stored as mesh, depth and vision value. The first index of finger's array represents thumb to little finger. The second index of finger's array represents the joint of the finger. Palm value recognizes and stores when the user touches the object with finger and palm. The stored vision images and each value are used for recognizing object as well.

$$\text{Finger}[i][j]\text{-Threshold} \le \text{Finger}[i][j] \le \text{Finger}[i][j]\text{+Threshold} (i \neq 0) \quad (1)$$

$$D_{min}[i][j] \le D_{Finger[i][j]} \times 0.95 \le D_{max}[i][j] \ (i \ge 0) \quad (2)$$

Finally, the system provides learned gesture recognition and detects defined interactions. These methods refer to the stored database (Table II). The system recognizes touch gesture when the user does the same gesture as the user-defined touch gesture. However, it is difficult to touch exactly the same position as defined. Therefore, we implemented the recognition formula (see formula (1)). The system tracks the touched position from thumb to little finger except thumb. Each touched finger between the thresholds (formula (1)) can be detected. An appropriate value for threshold is 0.3 for each depth value (x,y,z axis value). It had good result in all experiments. After detecting learned touch recognition, the system tracks the user's gesture and records mesh and point cloud motion. The touch gesture motion and interaction can be paired by user. The interaction type is determined as well.

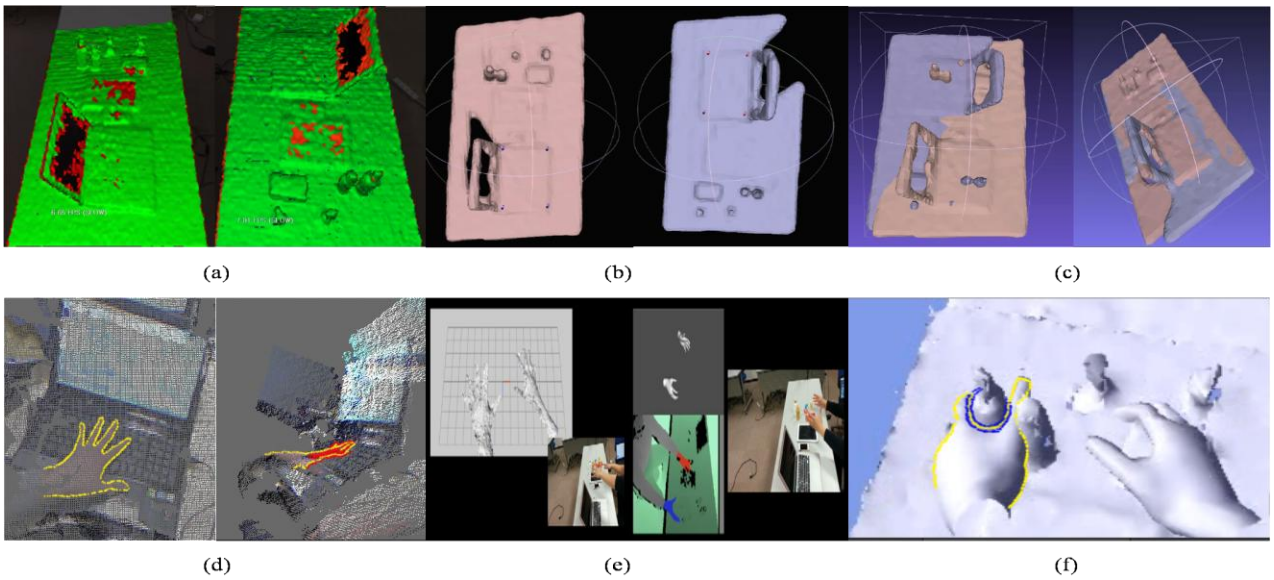


Figure 6. (a) 3D scanning using depth image from RGB-D camera1 and camera2 (b) Point cloud based reconstructing objects from two cameras (c) Calibrating two different direction RGB-Depth data and hole filling processing (d) Hand touch recognition with calibrated point cloud data (e) Hand shape learning and modeling (f) Grasp touch recognition using point cloud based 3D modeling estimation

The interaction returns feedback when the user does touch gesture using these databases.

C. Touch Recognition Technique

The touch recognition is conducted using point cloud data and 3D model estimation. Each finger is divided into three parts such as the finger joint for robust touch recognition (see Figure 7b). Each finger joints have 3D point cloud data. Our system tracks the location of the touch between the joint and the object in each joint. It then calculates the number of touched point cloud data and their motion for each joint. Next, it determines touch recognition based on formula that we found empirically (see formula (2)). The system determines whether 95% of the finger joint cloud data are touched in the same position that the user defined or not for each finger joints. Our system can estimate the area that cannot be seen by the camera because the system knows D_{min} and D_{max} value of each finger joint. The finger point cloud data between D_{min} and D_{max} is considered touched even if cameras cannot see the entire area. The system can recognize a 3D touched area with an object using this proposed method. The error rate of recognition is low because each finger joint is managed separately. The system can determine using three finger joints data even if an one finger joint cannot be well recognized.

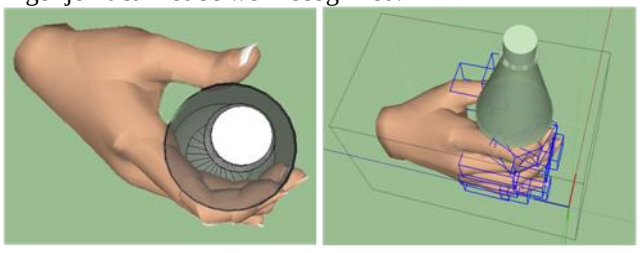


Figure 7. (a) Grasp touch object (b) Dividing finger joint of each fingers

The system uses middle joint of each finger mainly when all the three finger joints touch with object as well. Because the system can estimates the first and third joints point cloud data using middle joints data when all finger joints touched (see Figure 8b).

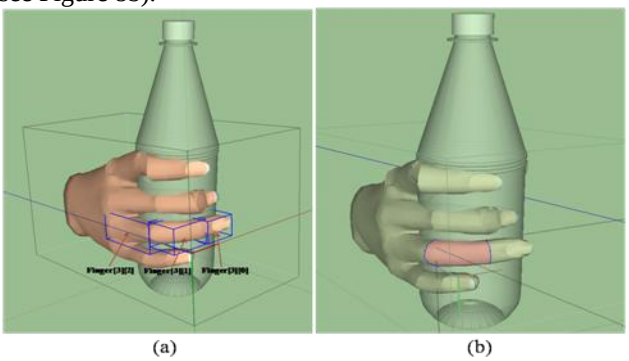


Figure 8. (a) Finger joint of ring finger (b) Point cloud data of second joint in ring finger

D. Tangible Interactions

After pairing objects, the system can be used in various tangible interactions. We described tangible interactions using the proposed touch recognition technique. We paired the interaction functions to the learned touch gesture using touch pairing. We implemented five types of interactions; flight game, map control, music control, instrument and drawing using existing object touching gestures. We describe in detail their functions and gestures.

1) Flight game and map control interaction: We designed a flight joystick using analog objects. Such a joystick can be used when playing a flight or shooting game. Our system made a pairing between a notebook PC and a lotion bottle with a cap as a game controller. The game can

be controlled by pushing the lotion bottle's cap (see Figure 9b) and moving it. The movements can be recognized based on the cloud data for the object and the hand. We also designed a map controller with paired objects (see Figures 5d and 9c). Our system tracks the two toys and the touched position; the map is moved to provide feedback.

2) Music player interaction: We implemented a simple interaction: a volume controller using toy. The system stored the toy object by calibrating and reconstructing. The user performed clockwise or counterclockwise rotations for learning gesture. The system stores touched location of object and hand motion of the gesture to the database. The user makes connection between the gesture and volume up or down interactions. After that, the music player's volume is changed when the user does a clockwise or a counter clockwise touch gesture (see Figure 9d).

3) Instrument interaction: We implemented an instrument controller using existing objects. The objects represent parts of drum instruments in this interaction. The multiple objects touch recognitions are conducted for doing learning gesture. These recognitions are used not only depth and mesh values of objects but also vision data of objects, because it has to discriminate objects and multiple touch recognition with those objects. There are eight instruments in the drums on the tablet (see Figure 9e). Each existing object is defined to control two instruments. The user assigns touch gesture to the object. This interaction can control the volume and play styles by using learned gesture.

4) Drawing interaction: A drawing interaction is provided. The interaction is performed on the existing pen object as input device. Our system tracks the pen's movement to draw the interaction feedback. The drawing is performed when the user does writing action on the desk. We can write the character by making a writing action. The proposed interaction can change the pen color and control the thickness by touching the pen (see Figure 9f). The system obtains color changing feedback when the user draws after touching a user-defined location on the pen. We can extend to adding more function using touch recognition technique. It can be used as real tablet pen.

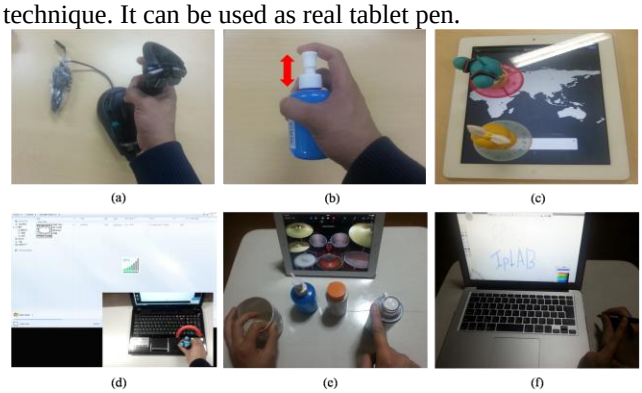


Figure 9. (a)Flight game joystick (b) Paired object controller as flight game joystic (c) Map control (d) Music player interaction (e) Drum instrument interaction (f) Drawing interaction

V. EVALUATION

We evaluated the touch pair system focusing on recognition accuracy. We evaluated touch recognition while changing the analog and digital objects alternately. We also evaluated 3D object recognition and tracking to demonstrate the system's usability and robustness.

A. Pairing Recognition Accuracy Experiment

Four analog objects and three digital objects were used. We evaluated finger touching, hand touching, grasping, and object-object touching for each object. The experiments were performed on computers (Intel Core i7 CPU, 2.5 GHz, and 8.0 Gb RAM) using two Microsoft Kinect sensors for Xbox.

We performed the experiments with 10 volunteers. We explained each touch pairing method. Then every volunteer performed four touching gestures to each object 100 times. We defined three second as the period for completing pairing via touching. When the system recognized a pairing, it showed color feedback. Touch recognition success alone was not counted. The participants were allowed to touch analog objects only with their fingers and digital objects only with their hand. Table IV shows the average pairing recognition success for the 10 volunteers.

Our proposed system showed >90% pairing recognition with the objects provided. We found that the average finger-based pairing recognition rate was higher than hand-based pairing. Finger pairing was recognized best when one or two fingers were used. Hand pairing requires checking whether the palm is touching. Thus, hand-based pairing recognition was less accurate than finger-based pairing. In addition, the success rate for touching a smartphone was lower than that for touching the other objects. This may be due to the size of the object. Most adult hands are bigger than most smart phones. Thus, it becomes difficult for the system to find the positions of the fingers and palm.

TABLE IV. FINGER AND HAND TOUCH PAIRING RECOGNITION RESULT

Digital / Analog	**Note PC**		**Tablet**		**SmartPhone**	
	finger	***Hand***	***finger***	***Hand***	***finger***	***hand***
Toy	98.3	95.4	99.1	95.3	95.3	93.2
Black Doll	95.7	93.3	96.2	94	92.2	90.1
Green Cube	98.4	96.7	99.3	95.7	96	94
Pet Bottle	96.7	95.3	97.1	95.4	93	91

percentage of recognition rate(%)

Table V shows the results for other pairing methods, such as grasping and object-object pairing recognition. The experiments were performed in the same way as described in Table IV. Grasping-based pairing recognition accuracy was >85% with the objects provided. This method uses point data from many directions. Generally, the front, side, and back surfaces of an object are touched when holding something with the hand. Thus, grasping has to be determined by analyzing the data from many directions.

TABLE V. GRASPING AND OBJECT-OBJECT TOUCH PAIRING RECOGNITION RESULT

Digital / Analog	Note PC		Tablet		SmartPhone	
	grasp	*object*	*grasp*	*object*	*grasp*	*object*
Toy	91.4	94.3	89.1	98.1	87.3	94.2
Black Doll	85.4	95.4	85.1	97	85.2	91.1
Green Cube	91.2	96.2	90.1	96.7	87	96
Pet Bottle	90.1	97.1	90.2	97.4	88	94.3

percentage of recognition rate(%)

Thus, as a whole, the recognition rate was lower than Table IV. We also found that the recognition rate differed by object size and hardness. The plastic toy, green cube, and bottle are relatively hard. However, the doll is very soft. When the user touches or grasps a softer object, the system encounters some difficulties in determining the touch depth. Thus, recognition accuracy was lowest for the doll among all objects provided. However, generally, the touch recognition rate was high enough to be uses for touch pairing system and tangible interactions.

B. 3D Object Recognition Accuracy

1) Single Object Recognition: In this experiment, we present the results for 3D object recognition with an 3D object database. Eighteen existing objects were used for evaluation. We divide by hardness, into 2 group. Each group was divided into three group again by object size. The size based group includes three objects in total. The hardness value 1 means soft objects and value 2 means hard objects. The size value 1 means small size such as pen, eraser and card and value 2 means normal size such as beverage can, cup and plastic lotion bottle. 3 means big size such as cap, shoes, books, etc. We do not need to consider very big object because our system considered object that can be put on a desk (see Figure 1). The experiments were performed on a computer with two Intel Core i7 CPU 2.5GHz and 8.0 GB RAM computers and two Microsoft Kinect for Xbox 360. We performed the experiments with ten volunteers. We explained each touch method thoroughly. We touch the object by each touch method and stored 3D object database to all object. The place of object where volunteers touched is indicated by a sticker. The sticker shows the place to be touched by the volunteers in recognition experiment. A volunteer performed 10 times three touching methods to each object. We checked whether system recognizes touching or not. The graph shows the average of the number of successful recognition of every object from ten volunteers.

In result, we can obtain good recognition result in Figure 10. We can see that the result is lower in size 1 in comparison with another size. The grasp touch is lower in each size as well. This is due to the fact that a small sized object cannot be seen by the camera from all the angles. The area that cannot be seen is estimated. The grasp touch recognition is more complicated compared with other touch recognition as well, because the system has to consider more than two sides when grasp touch is occurring.

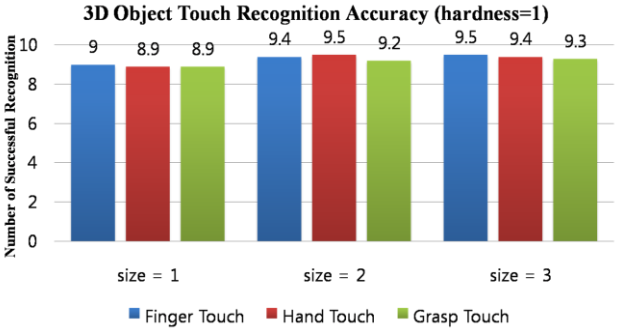


Figure 10. Touch Recognition Accuracy in hardness one

We obtained a higher result in Figure 11 than Figure 10. All size and touch methods scored over 90% successful recognition results, as seen in Figure 11. The recognition rate of hard objects is higher than soft objects, because the depth value was a little changed when the volunteer touched soft objects. As a result, the hard and big size object were obtained almost highest recognition rate.

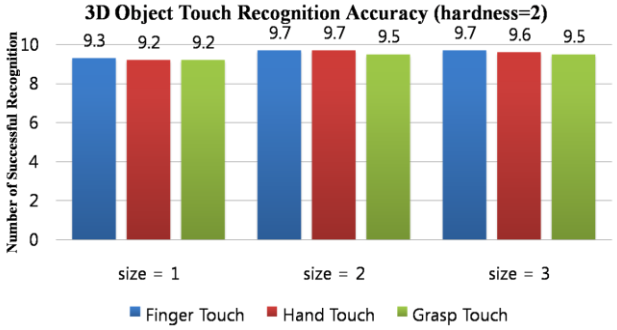


Figure 11. Touch Recognition Accuracy in hardness two

2) Multiple Object Recognition: The volunteers choose the objects that they want to use. The object size and hardness was not considered for natural situation, because the size and hardness of objects are not considered so much when we use existing objects in real life. In this experiment, we have already stored all objects in 3D object database and position that are indicated by a small sticker. They choose two objects at first. After that, they touch one object of them. The system checked whether it recognizes touching or not and the correct object id (see Table I). Next, they choose two object that they want to use. After that, they do same way. The objects number are increased from three to five and repeated the same experiment. The experiments were done ten times for each group of object. The result graph shows the average number of successes with different touch method from ten volunteers. The obtained result was over 90% for all situations (see Figure 12). The result shows that the number of objects did not have an effect in the recognition accuracy. We limit the number of objects to

five because of the size of the desk in this experiment. However, the recognition rate does not get lowered so much even if the number of object is increased. The object size and hardness have an effect to the recognition accuracy more than object number.

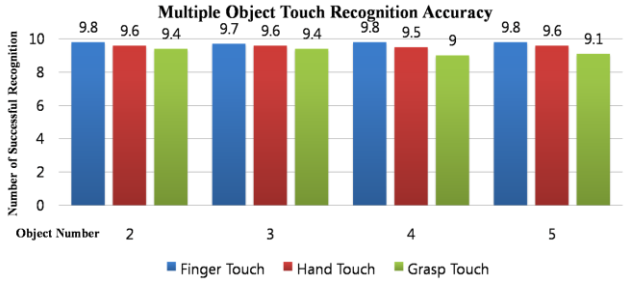


Figure 12. Multiple Object Touch Recognition Accuracy

3) Real-time 3D Object Tracking Accuracy: We evaluated object tracking after pairing. We moved analog objects during a 10 minutes experimental period (e.g., left-right, front-back).

Figure 13 shows the recognition accuracy for these tests. We obtained recognition rates of >80% for all objects. Since we used two Kinect sensors for real-time 3D object reconstruction and data comparison, we obtained low error rates.

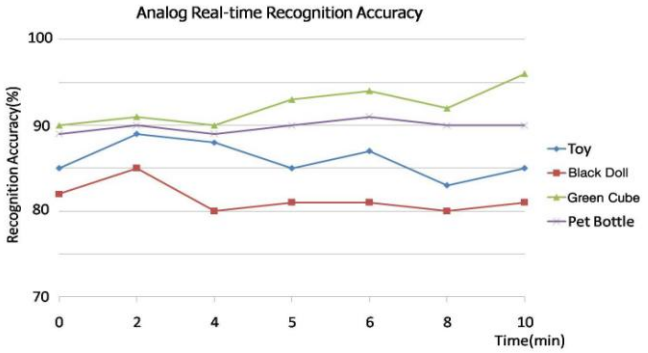


Figure 13. Analog Objects Tracking Accuracy

Figure 14 shows the recognition accuracy for three digital objects over 10 minutes. Smart phone recognition was <80% in around 4 minutes and 8 minutes after tracking.

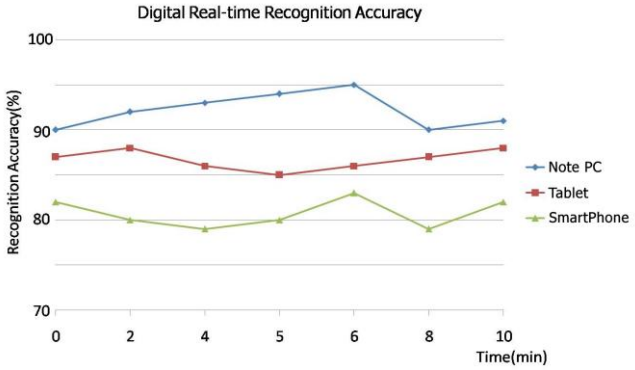


Figure 14. Digital Objects Tracking Accuracy

This is because the user was holding the smart phone with his/her hand. In particular, when we moved the smart phone with a front-back motion, the recognition rate decreased.

C. Learned Gesture and Tangible Interaction Recognition

In this experiment, we evaluated touch gesture recognition accuracy for tangible interaction using learned touch gesture database (see Table II). We defined the touch gesture and the system learned the touch gesture to store it in the database. We implemented five interactions such as flight game, map controller, music controller, instrument controller and drawing controller. We stored several gestures to each interaction. The flight game interaction was push gesture as shooting missile, moving gesture as moving flight. The music controller was clockwise as volume up, counterclockwise as volume down and double tap as play or pause. The map controller was same gesture with music player controller as moving map, zoom in or zoom out. The instrument controller was tap as play each instrument in the drum set, grip gesture as changing sound effect and swipe top from the bottom edge or down from the top edge as volume control. The drawing interaction was grip pen gesture in drawing mode, swipe down from top as pen's thickness changing and double tap for color changing. The defined gestures are evaluated by ten volunteers. They performed each gesture 10 times in the interaction. We checked whether the system recognizes the gestures or not. After that, we checked whether the correct feedback happened or not. The graph shows the average number of successes for each gesture in the interaction from ten volunteers. We obtained over 90% successful result (see Figure 15). The tap and double tap gestures to object are perfectly recognized. The drawing interaction and instrument interaction have similar gestures. The drawing interaction obtained lower rate than instrument interaction. The result was due to the object size, because it was difficult to capture all touched locations and motion of the pen object in drawing interaction.

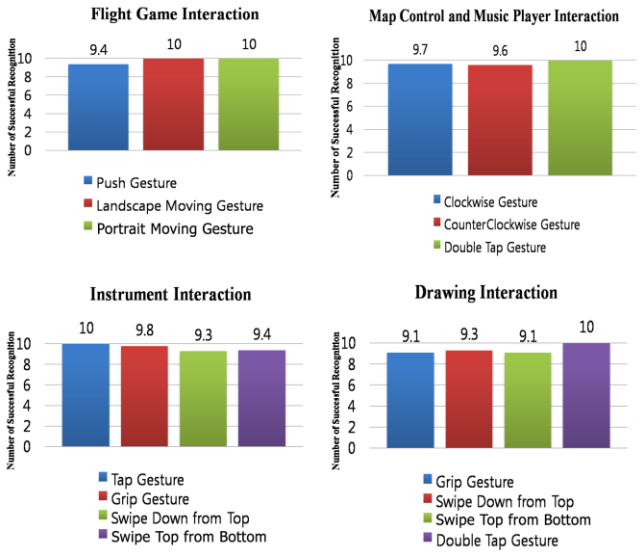


Figure 15. Recognition Accuracy of Each Interaction Gestures

We found that the object size and hardness are important for touch and gesture recognition in our system.

D. Discussion

In our experiments, we found out that the touch and gesture recognition with 3D objects can have good recognition accuracy rate. Our proposed method recognized 3D object touched location and learned gesture using RGB-D Kinect cameras without attaching additional sensors. We obtained good results over 90% successful recognition in all experiments. We found that the size and hardness of objects have an effect on the experiment results. The number of objects had no effect on the experiment results. The big sized and hard object's area can be captured for better touch recognition. The recognition rate is high because it is not an estimation. The estimated area of object for touch recognition is increased when the object is small or soft. In this case, the recognition rate is lower than the recognition rate for a big or hard object. However, the results between each experiment do not need to consider so much for using natural tangible interactions. The proposed system can cover their interactions and gestures.

In real life, we expected our natural actions to be used as interactions in general. For instance, a feedback happens when we grip the cup. The interaction occurs when we tap objects once or twice. A feedback is generated when we do clockwise or counterclockwise gestures with objects as real controller. It is interesting and useful when actions that used in real life are used.

The proposed interactions are designed by taking these points into consideration. The natural gestures are connected with interactions. The music controller, instrument controller and drawing interaction are functions that are frequently used. The designed gestures are friendly and frequently used as well. Therefore, the volunteers who participated to evaluation practiced well and became experts in using the proposed gesture and functions in learned gesture experiments. As a result, we obtain good results and proved the system's usability and robustness.

VI. CONCLUSION AND FUTURE WORK

In this paper, we described a new touch recognition techniques and dynamic pairing method using two Kinect camera based on 3D point cloud data. We described the touch pattern, touch system architecture and databases for recognition. We implemented tangible interactions using learned gestures that are stored by users with the proposed techniques. We did eleven types of experiments in finger, hand and grasping touch to evaluate our proposed system usability and prove the recognition robustness.

The accuracies were >90% for finger and hand touching and >85% for grasping and object-object touching. Almost the same results were obtained when we changed the locations of pairs dynamically. We obtained good real-time 3D object tracking results as well, despite using objects of different size, shape, and hardness. The accuracies were >89% for single object recognition in hardness value 1 and >92% in hardness value 2. We obtained the accuracies were >90% for multiple object recognition and >91% for learned gesture recognition. Especially, we obtained good real-time tangible interaction with good gesture recognition results in many situations as well, despite using objects of different size, shape and hardness. In addition, a remarkable result is the ability to implement tangible interactions with functions and frequently used gesture without installing any additional sensors to the multiple object. The contributions of the present study can be summarized as follows.

First, we have provided a new pairing method. Using this method, we can dynamically pair analog and digital objects, and various tangible interactions can be achieved.

Second, we presented a new touch recognition technique. There are many researches that recognize touching and motion gesture. However, they need to attach additional sensor to the objects and cannot recognize well touched location of object from users. They cannot recognize natural gesture with touching existing object as well. By proposing this recognition method between objects, the various tangible interactions can be extended. Third, it can recognize multiple object touching and touching gestures. We can dynamically change the object that we want to be tangible object in real time. It means the proposed technique is natural and robust when we use it in real life, because we used multiple objects and gestures that we are used to in everyday life. We do not need to learn gestures or make sensor based object for tangible interaction. Our actions and existing objects are enough, and can be used for tangible interaction simply using proposed techniques.

Finally, in the experiments, we obtained good 3D object recognition rate and learned gesture recognition results. We evaluated three pattern of touch and object type such as by size or hardness for natural using and robustness in real time recognition. The learned touch gestures were evaluated with practical interactions such as flight game, map controller, music player, instrument controller and drawing. The users can expect practical and natural use of existing object as tangible objects because we obtained good results in recognition experiments. In near future work, we can expect various practical tangible interactions by using our proposed touch recognition system. The augmented graphical support interface and interaction can be implemented with existing objects. The existing objects can be remote controllers to control devices that are located at home using touch recognition system as well. For instance, the cup can be television controller, and then changed to light controller. The combination of interface with head mounted display is widely used; natural and interesting touch interaction can be expected.

REFERENCES

[1] U. Lee and J. Tanaka, “TouchPair : Dynamic Analog-Digital Object Pairing for Tangible Interaction using 3D Point Cloud Data,” in ACHI’14 The Seventh International Conference on Advances in Computer-Human Interactions, Mar. 2014, pp. 166-171.

[2] E. S. Martinusseon, J. Knutsen, and T. Arnall, “Bowl: token-based media for children,” in DUX '07 Proceedings of the 2007 conference on Designing for User eXperiences, Nov. 2007, Article No. 17.

[3] R. Wimmer and S. Boring, “HandSense: discriminating different ways of grasping and holding a tangible user interace,” in TEI '09 Proceedings of the 3rd International Conference on Tangible and Embedded Interaction, Feb. 2009, pp. 359-362.

[4] R. Wimmer, “FlyEye: Grasp-Sensitive Surfaces Using Optical Fiber,” in TEI '10 Proceedings of the fourth international conference on Tangible, embedded, and embodied interaction, Jan. 2010, pp. 245-248.

[5] OpenNI 3D sensing framework : Last visited on Nov, 2013. http://www.openni.org.

[6] openFrameworks : Last visited on Aug, 2013. http://www.openframeworks.cc.

[7] A. D. Wilson, “Using a depth camera as a touch sensor,” in ITS '10 ACM International Conference on Interactive Tabletops and Surfaces, Nov. 2010, pp. 69-72.

[8] P. Wellner, ″The digitaldesk calculator: Tangible manipulation on a desk top display,” in 4th annual ACM symposium on User interface software and technology UIST, Nov. 1991, pp. 27-33.

[9] I. Siio and Y. Mima, “Iconstickers: converting computer icons into real paper icon,” in 8th International Conference on Human-Computer Interaction: Ergonomics and User Interfaces, 1999, pp. 271-275.

[10] P. Ljungstrand, J. Redstrom, and L. E. Holmquist, “Webstickers: using physical tokens to access, manage and share bookmarks to the web,” in Designingaugmented reality environments DARE 2000, Apr. 2000, pp. 23-31.

[11] T. Nishi, Y. Sato, and H. Koike, “Interactive object registration and recognition for augmented desk interface,” in IFIP conference on human-computer interface Interact 2001, Mar. 2001, pp. 240-246.

[12] F. Klompmaker, K. Nebe, and A. Fast, “dSensingNI: a framework for advanced tangible interaction using a depth camera,” in TEI '12 Proceedings of the Sixth International Conference on Tangible, Embedded and Embodied Interaction, Feb. 2012, pp. 217-224.

[13] S. Azenkot, and S. Zhai, "Touch behavior with different postures on soft smartphone keyboards," in MobileHCI ’12, Sep. 2012, pp. 251-260.

[14] L. Cheng, F. Hsiao, Y. Liu, and M. Chen, "iRotate grasp: automatic screen rotation based on grasp of mobile devices," in UIST ’12 Adjunct, Oct. 2012, pp. 15-16.

[15] M. Goel, J. Wobbrock, and S. Patel, "GripSense: using built-in sensors to detect hand posture and pressure on commodity mobile phones," in UIST ’12, Oct. 2012, pp. 545-554.

[16] A. D. Wilson, "Depth sensing video cameras for 3D tangible tabletop interaction," in Proc. IEEE Tabletop 2007, Oct. 2007, pp. 201-204.

[17] A. D. Wilson, "Using a Depth Camera as a Touch Sensor," in ITS '10, Nov. 2010, pp. 69-72.

[18] A. D. Wilson, "PlayAnywhere: a compact interactive tabletop projection-vision system," in Proc. ACM UIST 2005, Oct. 2005, pp. 83-92.

[19] O. Hilliges, S. Izadi, A. D. Wilson, S. Hodges, A. Garcia-Mendoza, and A. Butz, "Interactions in the air: adding further depth to interactive tabletops," in Proc. ACM UIST2009, Oct. 2009, pp. 139-148.

[20] P. Dietz, and D. Leigh, "DiamondTouch: a multi-user touch technology," in Proc. ACM UIST 2001, Nov. 2001, pp. 219-226.

[21] C. T. Dang, M. Straub, and E. André, "Hand distinction for multi-touch tabletop interaction," in Proc. ACM ITS 2009, Nov. 2009, pp. 101-108.

[22] H. Benko, and A. D. Wilson, "DepthTouch: using depth sensing camera to enable freehand interactions on and above the interactive surface," in Poster Presentation at the IEEE on Tabletops and Interactive Surfaces '08, Mar. 2008.

[23] A. Agarwal, S. Izadi, M. Chandraker, and A. Blake, "High precision multi-touch sensing on surfaces using overhead cameras," in Proc. IEEE Tabletop 2007, Oct. 2007, pp. 197-200.

[24] A. Butler, S. Izadi, and S. Hodges, "Sidesight: multi-touch interaction around small devices," in Proceedings of UIST ’08, Oct. 2008, pp. 201-204.

[25] K. Kim, W. Chang, S. Cho, J. Shim, H. Lee, J. Park, Y. Lee, and S. Kim, "Hand Grip Pattern Recognition for Mobile User Interfaces," in Proceedings of the National Conference on Artificial Intelligence, 2006, pp. 1789-1794.

[26] P. G. Kry, and D. K. Pai, "Grasp recognition and manipulation with the tango," in 10th International Symposium on Experimental Robotics 2006, Jul. 2006, pp. 551-559.

[27] H. Ishii, and B. Ullmer, "Tangible bits: towards seamless interfaces between people, bits and atoms," in SIGCHI conference on Human factors in computing systems, Mar. 1997, pp. 234-241.

[28] P. Mistry, T. Kuroki, and C. Chang, "Tapuma: tangible public map for information acquirement through the things we carry," in Proceedings of the 1st international conference on Ambient media and systems, Ambi-Sys '08, Feb. 2008, pp. 12:1-12:5.

[29] H. Song, H. Benko, F. Guimbreti`ere, S. Izadi, X. Cao, and K. Hinckley, "Grips and gestures on a multi-touch pen," in CHI′11, May. 2011, pp. 1323-1332.

[30] S. Hwang, A. Bianchi, and K. Wohn, "MicPen:pressure-sensitive pen interaction using microphone with standard touchscreen," in CHI EA ’12, May. 2012, pp. 1847-1852.

[31] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon, "KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera," in UIST ’11, Oct. 2011, pp. 559-568.

[32] P. Lopes, R. Jota, and J.A. Jorge, "Augmenting touch interaction through acoustic sensing," in ITS ’11, Nov. 2011, pp. 53-56.

[33] A. Wilson, "TouchLight: An Imaging Touch Screen and Display for Gesture-Based Interaction," in ICMI 2004, Oct. 2004, pp. 69-76.

[34] A. Wilson, "Robust Computer Vision-Based Detection of Pinching for One and Two-Handed Gesture Input," in UIST 2006, Oct. 2006, pp. 255-258.

[35] B. Piper, C. Ratti, and H. Ishii, "Illuminating Clay: A 3-D Tangible Interface for Land-scape Analysis," in CHI 2002, Apr. 2002, pp. 355-362.

Explorative Design as an Approach to Understanding Social Online Learning Tools

Naemi Luckner and Peter Purgathofer

Vienna University of Technology
Institute of Design and Assessment of Technology
Argentinierstr.8/178
1040 Vienna
Email: {naemi, purg}@igw.tuwien.ac.at

***Abstract*—The everyday availability and use of technology has changed education as much as it has changed everything else. For 8 years now, we have used technological interventions to change a setting where we teach up to 800 participants per semester in a class, in order to make it more interactive, engaging, and interesting for the students. We document a snapshot from this ongoing process. Aurora is an online system that has been developed from simple experiments with existing tools and software to bring interaction to the crowd of learners. Over the years, it has turned into a solid and extensive collection of tools for online teaching, learning, and communication. This article traces the development of Aurora over two consecutive years. We document the structure of the system we developed, the insights from an academic year of using it, changes designed and implemented, and first evaluations of the use of the revised version.**

***Index Terms*—Asynchronous Interaction; E-Learning; E-Portfolio; Electronic Note Taking; Explorative Design**

I. Introduction

Introductory remark: This article is an extended and substantially revised version of [1].

The Web 2.0 [2] changed online culture and transformed it from a passive consumer culture to a participatory culture. This shift also influenced the process of teaching and learning, which is since referred to as E-Learning 2.0. The notion of E-Learning 2.0 is that Web 2.0 technologies are adapted and integrated in E-Learning systems [3]. Knowledge can be created, shared, remixed and re-purposed by communities of practice [4]. Students are part of this process, collect sources and participate in the communities, by sharing their own ideas and findings. Brown and Adler [5] describe a new age of education, in which lifelong learning is not only needed but also supported by the participatory architecture of the Web 2.0. They speak of a new learning approach, '...characterized by a demand-pull rather than the traditional supply-push mode...' of obtaining knowledge. They emphasize the importance of social learning in the new online learning environment, pointing out, that the traditional teacher-student relationship is exchanged by a peer-based learning relationship.

Siemens [6] built on this change of learning culture and devised his theory of connectivism. He states that learning in the digital age is the self-driven process of building up networks of knowledge. Nodes in a network can be data sources, communities or people and are connected to the network with strong and weak links. Weak links are more interesting since they can open doors to new areas of knowledge, diversity and innovation. Siemens points out that the life-cycle of 'correct' facts is getting shorter, and new knowledge is created faster, so memorizing facts is not yielding desired results anymore [6]. More important is the 'Know-where', which describes where knowledge can be found quickly rather than learning the knowledge itself by heart.

In this paper, we present an E-Learning System with the aim of letting students take responsibility of their own learning process. The system is an attempt to create a holistic learning platform, valuing not only assigned course work, but also social interactions and additional content students create or discover over the course of a semester. We wanted to avoid to develop another system increasing the distance between teacher and students. Instead, our goal was to start from the rather difficult situation of very large classes, where contact between teacher and student is short and rare, and transform it so that students have a feeling of more immediate involvement, more contact, and more personal mentoring. To achieve this, we put concepts such as social interaction, participation, and exchange at the center of our design efforts.

Aurora is a learning platform that consists of three modules that can interconnect with each other. Firstly, the Dashboard is an administrative tool, containing an administrative Newsfeed as well as widgets to enhance communication between all participants of the course and maintain an overview of the course progress as well as interesting developments around the course topics. Secondly, the Slides module is used during and after lectures as backchannel and basis for upcoming discussions around course topics. Thirdly, students are provided with a pool of activities they can choose from in the Portfolio. We chose the word 'activity' rather than 'exercise' for work assignments, since we want to motivate students to actively pursue their work for this course and we want to avoid the vocabulary usually associated with course work to try to increase motivation. The Discuss module is used for discussions surrounding the topics that are covered in the courses. The name Aurora is not an acronym, nor has it any deeper meaning. We used the name because it refers to something beautiful, and because it sounds appealing.

The remainder of this paper is structured as follows: The next section 'Overall Goal' will give an idea of our motivation and process to create a new e-learning system as well as explain why we chose to develop the system ourselves instead of taking existing tools. Subsequently, in the section 'Components', each module of Aurora is described and compared with existing solutions from literature. The description is followed by a preliminary 'Evaluation' and 'Conclusion so far', in which we describe how the evaluated data influenced future designs. The section 'Iteration and Redesign: Another Version of Aurora' presents a new design of the e-learning platform *Aurora* that was used in the summer semester 2014. The biggest change of this redesign is discussed in the section 'Challenges', which is a new module replacing the Portfolio module. Finally, a section 'Future Plans' outlines upcoming iterations of the platform.

II. Overall Goal and Approach

At the Vienna University of Technology, lecture participation is sometimes in the high three-digit numbers. Traditionally, this would mean that lectures have to be endured with a passive and consuming stance. Around 2005, we set out to explore new ways to make lectures more interactive. We started by appropriating existing systems like IRC and Twitter to facilitate backchannel communication and interaction for students visiting large lectures. Early on, we were fascinated by the idea that we could time-sync this information to the slides. This would enable students to understand the backchannel as a means of taking (collaborative) course notes that became attached to the individual slides of the lecture.We also started to replace the then prevailing passive HTML web pages for course information with blogs, which seemed an ideal fit for some years.

As we better understood the necessities of the context, we faced two possible directions for the further development: use existing systems and services to piece together a larger system, or implement a whole new system according to our needs and ideas. Comparing these two approaches, we found several advantages of the latter over the former. One advantage is that in a custom system, we could make sure that students get by with a single login, as compared to multiple logins in a setting where several existing services are stitched together. Also, if we build the system ourselves, we can experiment much more freely with the organisation, structuring and interaction of the system, compared to pre-existing solution. Finally, as we are part of an informatics faculty, this approach also gave us an opportunity to offer meaningful master theses projects to our students.

So, as we better understood the necessities of the situation, we supplanted the use of an existing blogging solution with a custom-made Newsfeed implementation that was heavily inspired by the structures and aesthetics of social media systems like Facebook or Twitter, still offering us more control over composition and access for us.

By and by, we replaced all passive elements of the information infrastructure for our large scale courses with interactive components, we also set out to change the way we evaluate student performance in order to come to a final grade. This led to a somewhat idiosyncratic redefinition of a portfolio system that we implemented.

All these systems are currently being actively developed and refined in an effort to explore new ways of teaching and learning for a generation that grew up with ever-present Internet access and for the most part played a lot of games [7]. We redesign our systems year after year after understanding what works and what does not. We pursue this research in the spirit of design as research, or explorative design. One core idea is that with each version, new concepts become evident that were not yet visible last year, be it from use, from formal or informal evaluation, or because we reflect on our progress from the feedback we get from students.

As approaches such as participatory design, contextual inquiry and user involvement are deeply rooted within the institute this project is created in, we used multiple approaches to make sure the interests and perspectives of students are considered in the design process. Those approaches include:

- offering the users a continuous feedback channel that was constantly monitored by project members, making sure that all issues are responded to accordingly;
- offering opportunities for students to do bachelor or master projects within the project, exploring new directions and implementing novel ideas;
- organising user testing sessions during active development with students from previous semesters as testers;
- starting the semester with a week labelled as trial run, where feedback was especially appreciated;
- offering exercise activities framed within the content of the lectures using Aurora where students could reflect on the concept, features and design of the system, and propose redesign ideas;
- obtaining structured feedback using questionnaires widely distributed among the students;
- organising semi-structured feedback rounds at the end of each semester in order to talk to the participants about what worked well, what did not work so well, and what was missing. These sessions routinely turned into co-design sessions where new ideas were proposed, discussed and evaluated.

Following this path for some years now, we have come to a place where individual components have been published about, but we never set out to describe the system as a whole. This is what this paper sets out to do.

III. Components

Aurora is a collections of different components, each of which takes on a vital role for the lecture to run smoothly. There needs to be a place to publish information about how the lecture is run and how it is graded, a place for the lecture content and a place for work to be done by the students that is evaluated by staff. In the following sections, we describe each of the solutions we implemented for these requirements in detail. Each section is preceded by a literature review of

relevant other work in the same design space, in order to provide an overview of how others previously approached similar problems.

A. Dashboard

Dashboards are often used in complex system to provide participants with an overview of activity on these platforms. The role of a dashboard is variable, depending on the context of its application. Dashboards have been used to track activity from different applications in a complex system [8]; to create peripheral awareness, provide navigation, and a system-wide inbox [9]; to create awareness of group members' actions and to convey the status of shared artifacts [10]; and to provide multiple views of a large dataset in a system [11]. More specifically, in an e-learning context, dashboards have been used for self-monitoring for students and to improve teachers' awareness [12]; and to help students to relate their learning experience to that of their peers or other actors in the system [13].

In Aurora, the Dashboard is the first page every student is presented with when logging into the system. It is a collection of widgets, containing the Newsfeed, an individual course status overview, showing colleagues, groups, current links and additional contact information. The page draws together course-relevant information related to the content from other websites, as well as information from other components of Aurora.

In former versions of Aurora, we included a statistics page to enhance students' peripheral awareness. The page provided a statistical overview of the data that is distributed over the whole system. Students could, for example, look up who of their peers was involved in a lot of discussions, or who got a lot of stars, which could be awarded for good comments by other students and members of the staff. At the start, seeing an overview of the work done in the system had a motivating impact on students and staff alike. Especially dedicated students could easily be singled out and earned a good reputation and trust among their peers. However, it also created a ranking among the students, which changed a lot at the start, but after a while it was very hard to move up ranks, which had a negative effect on some students. Since we did not want to strengthen competitiveness in the course, we first decided to hide the statistics view from students, and, in later versions, the view has not even made it into the system because of a lack of time and resources for the development.

1) Newsfeed: The Newsfeed is a largely organizational message board, but can also be used for content related postings. The lecture staff can use the Newsfeed to publish course updates and other relevant news for the students. Questions, annotations, complaints and praise can also be posted here, and can be answered by other actors in the system. Students post content related comments as well, but are asked to first look for a suitable slide in the Slides section to provide context for the content, before blindly posting it in the Newsfeed.

Information from other components is collected and posted via sticky notes at the top of the Newsfeed. Students are informed if someone answered to one of their postings in the Slides section and can jump directly to the posting via link. If students get points for a good comment in the Slides section or for a newly marked activity in the Portfolio, they are notified here. Direct messages show up on top of the Newsfeed section, and can be sent by either colleagues or team members.

The Newsfeed enhances direct communication between students and staff and also provides a forum for discussions about the course design. It can be searched or filtered to see either only staff postings, only organizational postings, or only content related postings. Students can subscribe to Newsfeed postings via RSS to integrate them into their everyday online environment.

2) Additional widgets: The Progress Bar widget is a tool students can use to get an overview of their progress in each of their classes. Each lecture has an overview of the student's activity status. It shows the amount of points received in the lecture through activities and comments, as well as the total amount of points. Additionally, it shows how much work the student has handed in but that has not yet been graded, and the how much the student can still hand in until the end of the semester.

In the Colleagues widget, users can add other students to their course network and, on acceptance, see their avatars and further information. They can write direct messages to their colleagues as well as see all their colleagues' comments in the Newsfeed and the Slides highlighted. This intends to create a feeling of connectedness within the course and motivate to interact with others regularly.

Some activities in the Portfolio can be worked on in teams. The Teams widget shows a list of all existing teams the student is a part of. Each entry contains the name of the project the team is working on, the possibility to send a message to all team members, and a list of the other team members.

The Current Links widget displays a list of recent articles and interesting websites - supplementary reading material of topics covered in the course. The collection of links is compiled in a blog using soup.io and integrated into the Dashboard via RSS.

Furthermore, the Dashboard lists contact information to correspond with the staff directly. Students are invited to ask all course relevant questions directly in the Newsfeed so that other students can profit from the answers as well, but some issues (e.g., personal problems) need to be taken up with the staff directly.

B. Slides

There is some research on how to offer interactivity in large lectures. One approach are 'Audience Response Systems', also called 'Clickers'. Kumar and Rogers pioneered such systems in their 1976 'Olin Experimental Classroom' [14] that featured a feedback channel for students in the form of 12 buttons. Today, clickers are commercially developed products, offering a number of potential benefits to large lectures. Caldwell [15] summarized the literature on using clickers in lectures. Recently, software clickers have begun to appear, based on

the fact that most students bring a network-connected device, most prominently mobile phones, to lectures, but this approach is still mostly experimental [16] [17].

Of course, more elaborate backchannel communication systems have been tried as well, such as ActiveClass [18], Fragmented Social Mirror [19] or ClassCommons [20]. The development and evaluation of these systems overlaps with the development of the approaches presented here, first published in 2008 [21].

It can be argued that backchannel communication during lectures is potentially distracting, diverting the attention from the speaker to unrelated things. On the other hand, students regularly bring their laptops to class in the hope of finding productive use, but often end up being distracted by other things that are available on their computer. We have observed that supplying students with a backchannel that is centered around the lecture itself brings some of that attention back, and while it creates 'bubbles of diversion' from the lecture itself, at least these bubbles are focused on the content of the lecture.

Slides consist of two major components, Livecasting and Studio. Livecasting lets participants add notes to individual slides of a lecture, either in the style of a backchannel conversation, or privately. Once the lecture is finished, slides and comments are available in a combined view in the Studio. Participants can keep adding comments, links etc. in the Studio, so that the lecture slides become the focal point of discussion and exchange for participants and lecturers alike.

1) The Livecasting component: During a lecture, the lecturer runs a script on her computer. By pressing the next slide button on an accordingly conFig.d remote, she triggers a script that sends the number and title of the newly displayed slide to the Livecasting server (Fig. 1). Additionally, the script retrieves the lecture notes of this slide in the presentation document and scans them for a custom-made meta-syntax signifying information that is meant to be posted with the same slide. These text-lines include explanations, enhanced quotes, references and other links, activities and discussion starters.

Students load a web page that changes with each slide the lecturer shows, offering them fields to enter public comments and private notes. Information entered into either of these fields ends up being attached to the slide that was visible when the participant started typing. Like in a chat system all connected participants can see the public comments entered by other participants, and they can reply to these comments, creating ad-hoc discussions of the lecture content. To ease the cognitive load, a participant's own comments are colored yellow. Additionally, students have the opportunity to mark slides as 'liked', 'important' or 'unclear' with a single click.

2) The Slides Studio: Once the lecture is over, the lecturer makes slide-by-slide images of his presentation available to the Slides module (Fig. 2). While this can also be done before the lecture, we decided not to show the slides because of the obvious spikes in network traffic this would generate whenever a new slide is shown. The slides, all the participants' comments as well as the lecturers automatically posted comments are then made available in the Studio.

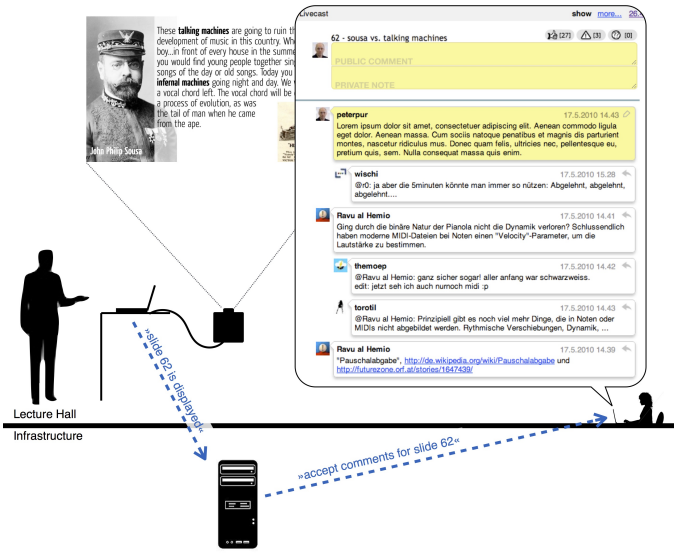


Fig. 1. Structure of the Livecasting setup. While the lecturer talks about a slide, connected students are offered text entry fields where they can attach public comments or leave private notes with the slide. Because public comments are immediately visible to all other students, this creates a setup similar to an instant messenger that is contextualized by the slide currently projected for everybody to see.

Here, participants and lecturers can post comments even after the lecture is finished. In the Studio, the slides are arranged horizontally, sorted by their time of appearance in the lecture. The comments attached to each slide are laid out vertically, with the earliest comments up on top (usually, these are the comments posted by the script on the lecturers computer immediately when the slide is shown), with reply threads sorted in the same way.

Participants can give praise to good comments by clicking the star next to the avatar of the author, in which case the star turns yellow and shows the number of clicks it has accumulated. Lecturer can use this same mechanism to award points to outstanding comments. In this case, the star is distinguished with a green glowing outline, making its commendation visible to everybody.

While lecturer's comments are generally displayed in the same way as student comments but distinguishable by a light-blue color, there are two lecturer-posted types that stand out from the rest: discussion starters and activities. Comments of both these types are arranged between the slide and the 'private notes' border, thus standing out even when scrolling through the slides quickly.

Discussion starter comments typically contain a questions and an invitation to discuss this question in the comments of the slide. We use this mechanism to initiate discourse on the content of slides worth of discussion, and to initiate discourse between lectures, asking participants to discuss upcoming content. Activities contain a brief explanation of an activity, linking into the Portfolio system where an elaborate description of this activity can be found. This gives the lecturer

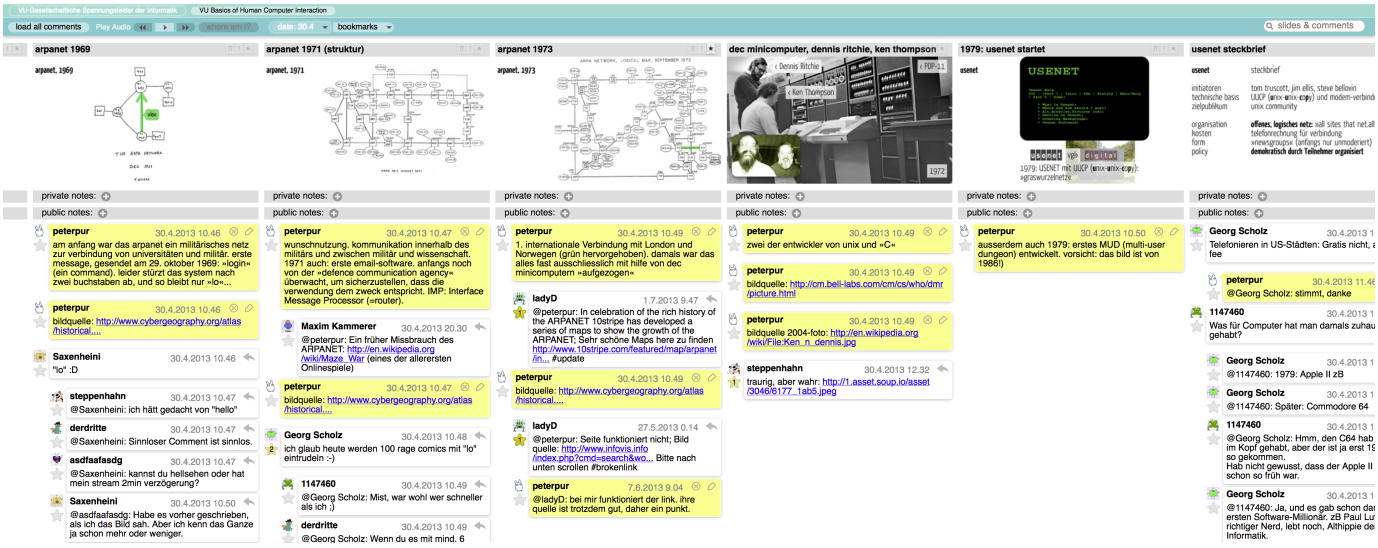


Fig. 2. Slides Studio, where all slides and comments become accessible to participants and lecturers alike. Note that your own notes are colored yellow.

an opportunity to announce new activities that derive from the content of a slide. Activities comments link students to the Portfolio of Aurora, where they hand in their work for review and evaluation.

C. Portfolio

In areas like HCI or Informatics and society, it is hard to conduct written exams, and once you have more than a couple of hundred students, it becomes impractical to the point of impossible to conduct oral exams. We started to abandon tests and exams at some point when we made the observation that the prospect of a written exam changed what we taught. This compromises the whole idea of teaching and learning a subject matter, especially at the university level.

For a couple of years now, research papers have been explaining the theoretical sense the adoption of ePortfolios would make. Advantages implied are, among others, 'improved reflection, increased student engagement, improved learning outcomes, and increased integration of knowledge' [22]. The paper quoted gives a comprehensive overview over ePortfolio research, and points out the lack of empirical support for many of the asserted advantages.

The module we call Portfolio is not really an ePortfolio in the strict sense of the word. While we explicitly ask the students to upload artifacts that show what they have learned, we offer a large catalog of predefined activities that can be handed in here (Fig. 3). These activities include a broad range of tasks, from simple applications of theoretical content, to actions reflecting their own prior projects, to complex design exercises. Many of those activities would make viable exercises in a traditional deadline-based context, while others would be quite unsuitable for such an environment. The catalog also contains meta-activities such as finding new sources, suggesting new activities, and organizing round table discussions with experts in the field. No single activity yields substantially more than 10% of the final grade, so that students will be exposed to a broad range of topics.

Additionally, we attached a commenting section to each of the activities, designated as 'Q&A'-area, where students can ask questions regarding the activity that will be answered by the course organisers.

Participants hand in their work using the portfolio system of Aurora. We do not set any specific deadlines other than the end of the semester, and we do not expect them to follow a specific order. The only requirement they have to meet is to make sure that their work is distributed throughout the semester, instead of congested at the end. To this end, we devised a system to keep students on track by pushing them to regularly hand in work over the semester.

Each students has to reach a certain amount of credits in order to successfully finish the course. Each activity is worth a given amount of credit points that, when handed in, count towards the final grade. During the semester, students manage a certain contingent of 'possible points' they can hand in at any given moment. The number of credit points of each activity handed in are subtracted from this contingent, leaving the student with only a small amount of points left to be handed in at that time. Each day though, the contingent is replenished a little, to the point where it is full. If the student does not hand in activities before the contingent is totally refilled, they start losing those points. On the other hand, as long as the student does not have the needed amount of points in their contingent, they cannot hand in new activities. That way, students have to continuously hand in activities over the semester, but can chose when to do so. They get enough points to easily finish the course, even if they loose some of those points along the way. This system basically makes sure that if students waited too long into the semester, they would be unable to accumulate enough credit to complete the course.

The Portfolio includes an easy-to-use review component for the course admins to review and evaluate the participants'

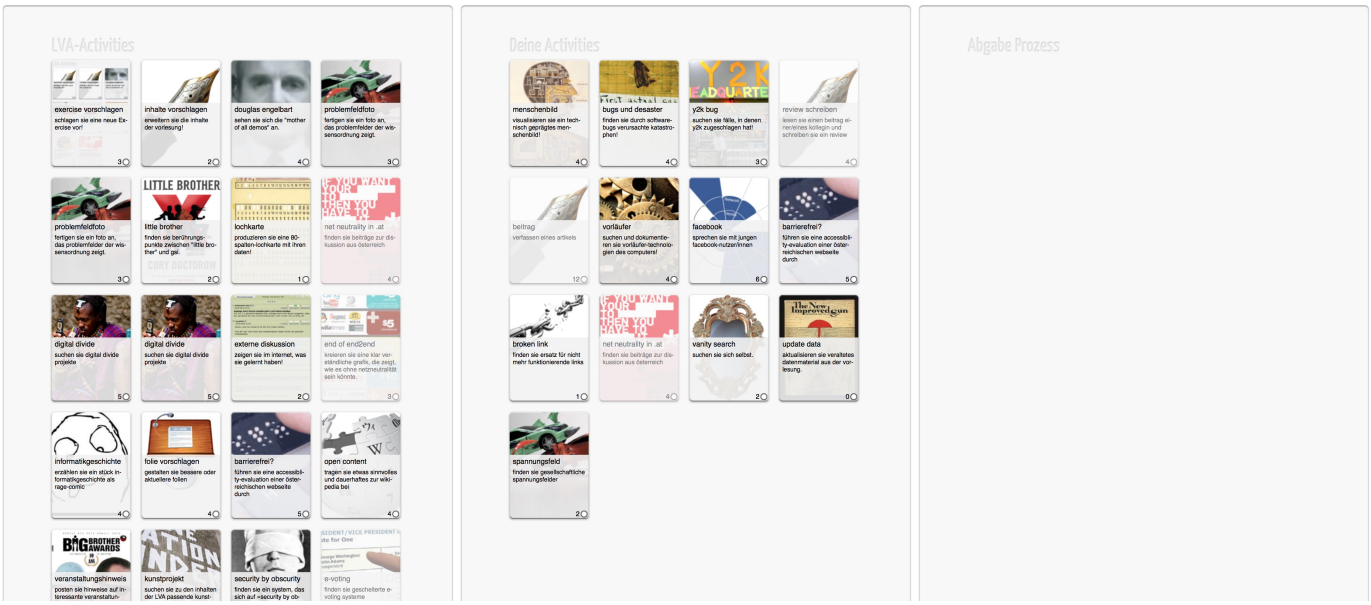


Fig. 3. Portfolio view of a participant, with the catalog of available activities, shown as cards, in the left column, a working area in the middle titled 'Deine Activities' (Your activities), and the area for hand-in in the right column. A student would drag an activity card from the catalog into the working area to elaborate, and later drop it into the hand-in column for evaluation.

work, with the notable addition of enabling students to hand in repeated submission of work that failed to meet the standards. It also includes a double blind peer review component that makes part of the assessment process into an activity by itself on the premise that if you do an honest review of somebody else's work, you will learn a lot. This functionality was used for an activity where students wrote in-depth articles of some more or less freely chosen course content, targeted at an imaginary journal. The articles were then double blind peer reviewed by other students.

The organisational approach described here tries to abandon the usual scattering of deadlines through the semester, giving the students a lot of autonomy in their work, which self-determination theory deems essential for intrinsic motivation [23].

D. Discourse

Discussion systems are widely used online, also in an educational context. Research shows that discussions foster active student participation and knowledge transfer [24], train critical thinking skills [25], and are used as a communication channel between students and teachers [26]. In the context of this project, a discussion system has been created, with an emphasis on redesigning threaded discussion systems to effortlessly join long discussions and easily follow single discussion threads. The effect of layout generally ([27], [28], [29]) and of layout in discussion systems ([30], [31]) has been discussed before.

In our opinion, traditional online discussion forum systems share a couple of common problems. For example, as the number of postings grows, readers lose track of all the places where they posted something. This often leads to users reducing their involvement in order to retain the feeling of control. Another problem is that intense discussions between individual participants can quickly derail a discussion, making useful and on-topic contributions hard to find. Problems like these seem to come from the way information is presented to the user, suggesting that a better visual structure and more adequate interactive organisation could improve on these problems. Thus, we set out to design and implement a completely new system for online debates.

Discourse is an asynchronous, multi-threaded discussion system, that can be used for on-topic discussions among students. Lecture staff can post discussion starters or additional material to a slide. Students can navigate to the discussion via a link, which opens an infinite discussion canvas, inspired by Scott McCloud's *Infinite Canvas* idea [32]. There, each discussion is displayed in two dimensions: vertically and horizontally. The vertical axis is used for new ideas, thought and inputs into the discussion. Each comment can be replied to, which in turn creates the second, horizontal dimension: replies are displayed in a new column to the right of the original reply. Only one thread of the discussion can be opened horizontally at any given time. Fig. 4 shows an example of an open discussion thread.

A more detailed description of this format as well as an elaborate evaluation of the effect of this specific layout and interaction on the content of the discussion can be found in [33].

IV. Evaluation

Our focus in evaluating these components is a better understanding how we can advance the system. We do not have an ultimate goal, but we use both the design process, in the sense of 'doing for the sake of knowing' [34], and the evaluation to understand how the system should be enhanced,

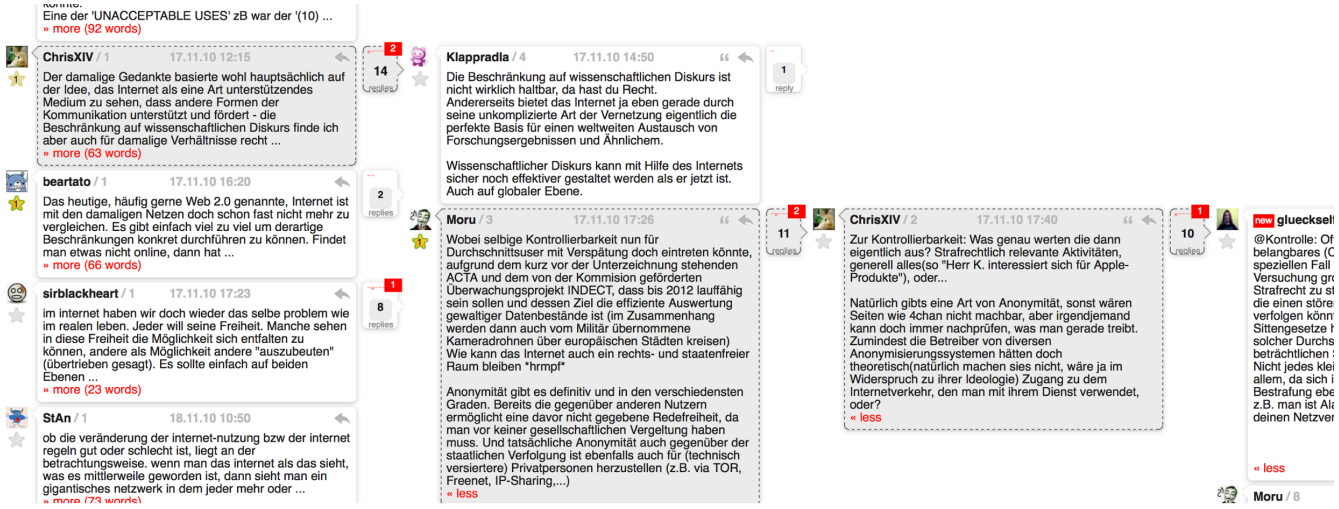


Fig. 4. A horizontal discussion thread. Selected items are displayed grey and with a dashed outline. Information boxes on the top right of each comment provide a quick overview of the rest of the discussion thread.

refined and changed in order to satisfy our needs as teachers as well as the needs of the students as learners. We understand that each step in this process changes the situation, leads all participants to new and often unforeseeable behaviour, which in turn influences and challenges all the assumptions we made to come to this point. This is why we refer to this process as explorative (or exploratory) design.

TABLE I. The table shows how many people were involved and how many certificates were handed out in the lecture. Note: columns do not add up because teachers and students could be associated with both courses.

	Profs	Predoc	Tutors	Students	Certificates
BHCI	3	1	6	733	442
IST	1	1	4	521	337
Total	3	1	10	842	779

This partial evaluation is based on data from two courses, Basics of Human Computer Interaction (BHCI) and Interactions of Society and Technology (IST), which took place in the summer semester of 2013. A total of 11.793 activities was handed in over the course of the semester, 7126 in BHCI and 4667 in IST. The staff of both courses combined consisted of 3 professors, 1 predoctoral fellow and 10 tutors, exact numbers can be found in Table I. Students only got a certificate if they handed in at least one activity. Every student who ultimately received a certificate handed in 15 activities on average. In the Slides section, 1283 slides were posted distributed over two courses with 23 lectures in total, and 3975 comments were written during and after these lectures.

A. *Portfolio evaluation*

Fig. 5 shows a pie chart of the time it took to grade activities. One third of the activities were graded after a week, which would be an acceptable amount of time for students to wait for feedback. Given the student-staff ratio, we tried to achieve a maximum waiting time of three weeks until every activity is graded. As can be seen in Fig. 5, we were not able to reach that goal, as only two thirds of handed in work was evaluated within the given time frame. The final third of the pie chart consists of activities that took 4 and more weeks to be graded. Considering the importance of feedback in order to keep students motivated and continuously working [35], 4+ weeks seems too long a time to hear back on one's work.

Fig. 5. Time it took to grade an exercise, calculated in weeks

We suspect that this fluctuation in delay can actually be explained by queue modeling in game theory. Activities tend to be handed in unequally distributed time, leading to an overload that causes congestion and that is then almost impossible to resolve within the given resources until the end of the semester.

B. *Slides evaluation*

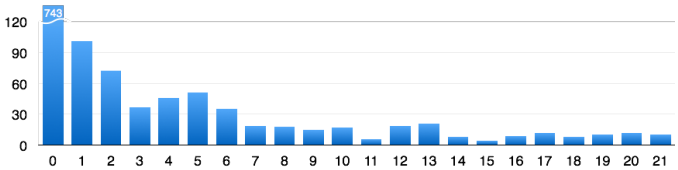


Fig. 6. How long after a slide was posted (Day 0) are students interacting with it via the comment stream. Note that the first column had to be shortened as indicated for reasons of scale.

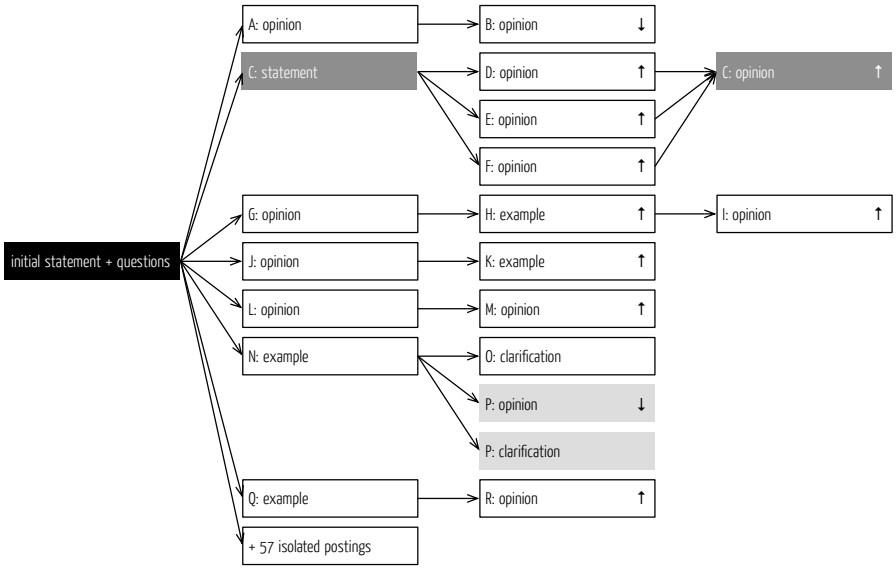


Fig. 7. The structure of a discussion in the traditional discussion, using the Slides Section

In Fig. 6, the comment data was analyzed to find out if and how long after the lecture students engage with the content by writing comments and discussing it in the Studio. The graph shows that most comments are written during the lecture, but there is a long tail (going up to 75 days) after the original slide was presented. Approximately 120 comments were posted even after the semester was over.

Especially interesting for us is the peak a couple of days later, as well as the 'long tail' of posted content after the lecture that can be seen in Fig. 6. An evaluation of these 'late' postings show that students came back to post information they find relevant, like news coverage, examples, references, etc. or to partake in discussions they have started with another postings. We see this as a successful feature of the system, as it induces reflection on and occupation with the content of a lecture for quite some time after the lecture is over.

C. Discourse evaluation

Two types of discussion systems have been used in two consecutive years; the first system used was the Slides Studio of Aurora, which features a traditional, one-level deep vertical representation of the comments. The seconds system used was the Discourse module. In each year, students were invited to join in voluntarily, and were rewarded credits towards their final grade for well written comments. For example comments that were argued conclusively and that featured links to additional material and sources.

We evaluated the outcome of 5 corresponding pairs of discussions on the same topic, one held in the vertical discussion section, the other in the 2-dimensional counterpart. For each of these discussions, we mapped the course of the debate as a tree structure, the results of which can be examined online at http://igw.tuwien.ac.at/designlehren/discourse_evaluation.html. One of these evaluations is presented here in this section, for more details, please refer to [33].

The discussion presented here was held on the topic of *phishing*. We analysed the structure using a qualitative content analysis based on Mayring [36]. Each posting was coded as either *Statement* (facts proved with sources), *Opinion*, *Question*, *Answer*, *Material*, *Example*, *Clarification*, or *Insight* (for details to these categories, please refer to [33]). Over all, most postings were identified as *Opinion*, which are position postings without sources. In the traditional threaded 'vertical' discussion format we also found postings of the type *Statement*, *Example* and *Clarification*. In Discourse on the other hand, discussions seemed to be more diverse in the types of postings that students wrote, much more engaging with more students contributing more than one posting, and much more in-depth as indicated by the fact that we found even comments coded as *Insight*, which are realizations of something that was so far unknown to them.

Fig. 7 shows the structure of the 1-dimensional threaded discussion. It was attended by 63 students and contained 77 postings, 57 of which were isolated postings. Only about 17% of the comments were written as a reply to someone else. Of the 63 participants, only 10 wrote more than 1 comment, and only 2 students ended up writing 3 comments each. The average amount of comments per students was 1.2, so most just wrote one isolated posting and left the discussion straight away.

The structure of same discussion in *Discourse* can be seen in Fig. 8. It was joined by 29 students and resulted in 50 comments, 19 of which were isolated postings. 50% of all comments created in the discussion were in reply to other comments. 8 students wrote more that 1 comment, 5 of them more than 3. Overall, each student wrote 1.7 comments.

The overall impression is, that *Discourse* leads to students being more involved in the discussion and even coming back to read up on new postings. The content seems more diverse

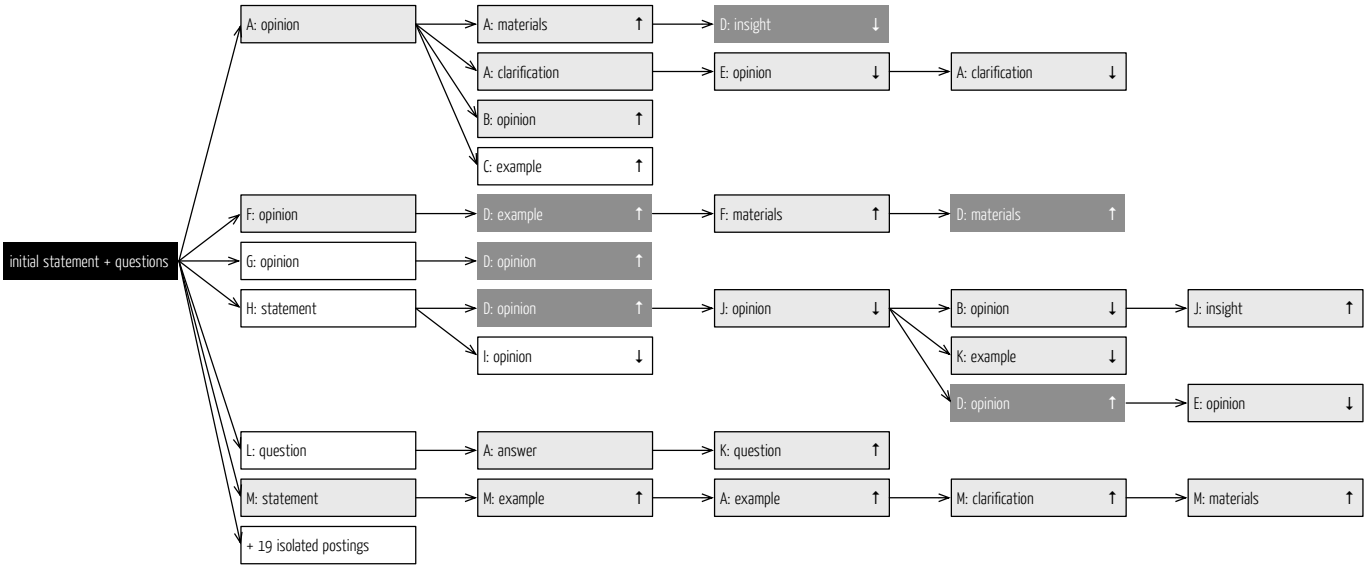


Fig. 8. The structure of a discussion using *Discourse*

and, in some cases, even signifies that learning occurred due to the discussion. The difference in the amount of isolated postings (19 vs. 57) shows that in Discourse, more students were motivated to find the correct discussion thread to post their own comment to, rather than just write isolated postings with no connection to what happened in the discussion until then.

V. Lessons learned

The main goal of our work is to explore the design space of online teaching and learning support systems. Our approach is best described as explorative design, with the main goal to better understand the context, the players, and their needs. At the same time, we acknowledge that technological interventions also transform the situation, and also, to a lesser extent, the needs of the players. In building and using systems that implement novel approaches to the context of teaching and learning, we in turn have a chance to understand the change such systems bring into the situation, and react accordingly. This approach shifts the focus of evaluation from understanding how and why the approach worked, or failed to work, to understanding and assessing the impact of an approach on a situation, and ultimately to finding new approaches to try. In the end, we are not so much focused on proving that our approach is right, e.g., by showing effectiveness by some abstract learning measurements. Instead, we want to find new and better ways to teach and learn that use the potentials of new technologies, engage and motivate students and tap their self-motivational capabilities.

The following conclusions were drawn from the use of the version of Aurora described in this paper so far:

A. Dashboard

The way the Newsfeed works to unify all organisational communication as well as general questions and discussions into one stream is promising. However, more effort has to be put into promoting important messages, which sometimes tend to get lost in the constant stream of incoming comments.

B. Slides

Slides demonstrates the potential to make content more interactive using novel forms of presentation. Discussions form around individual slides, students contribute additional resources and material, use the Slides module to pose questions and ask for clarification, and even share entertaining associations. The high granularity in the presentation makes it possible to post such contributions quite targeted, albeit for the price of overview. However, when a couple of hundred participants post comments, overview is hardly something one would expect to preserve.

A recurring critique of some students is the way information is organised in the Slides module. Specifically, those students who do not feel comfortable with horizontal scrolling, often due to constraints posed by their computer hardware, expressed a dissatisfaction with the principal structure of Slides Studio.

Also, the question prevails whether the slides used in the lecture sorted by date are the ideal organisational scaffold for such conversations. Slides often over-emphasise examples and illustrations, as those parts of the lecture often use more slides than abstraction, concepts and ideas, which are thus pushed into the background.

C. Portfolio

The use of a portfolio-based approach in the Portfolio module provided for flexibility and versatility unparalleled in prior version of Aurora, or compared to a traditional deadline-based course organisation. Many students appreciated the freedom and choice that come with such a system, while some students are clearly overwhelmed with the necessity to show such a

degree of self-organisation and self-discipline. Students did complain that the 'many small exercises' principle causes them to only superficially get in contact with a lot of interesting themes and questions, leaving them without a possibility to delve deeper and engage with some of the themes more in-depth.

At the same time, the tutors were flooded with a very large number (11.000+) of small exercises to evaluate, making it impossible to write explicit feedback and grade the hand-ins in a timely manner. One definite goal for the next version of Portfolio was to find an organisational form that enables us to keep the average waiting time for feedback within a week of hand-in.

On the other hand, it turned out that the 'Q&A' commenting section with each activity description was a advantage for the students. The amount of work generated is minimal compared to the value it has. For example, ambiguities in the description of an activity could be cleared up here quite efficiently and without generating version conflicts in the description.

D. Discourse

Discourse worked exceptionally well, resulting in more focused discussions, heightened participation and more interesting conclusions. Unfortunately, the student who implemented Discourse as part of his master thesis left at the end of the semester, leaving a system that worked reasonably well within the technical context of the old version of Aurora, but was very hard to incorporate into the substantially rewritten version that we used in the following year.

VI. Iteration and Redesign: Another Version of Aurora

After the evaluation of our experiences with the version of Aurora described in the paper so far, we set out to reorganise, redesign and implement a new version of Aurora. The resulting version has been developed during winter semester 2013 and used during summer semester 2014. In the following, we will briefly describe the changes planned, the redesigns taken and first results from the use of this new version.

A. Dashboard and Newsfeed

Based on last year's version, Newsfeed was improved in several places (Fig. 9). We added a dynamic filtering mechanism that lets user selectively shrink and expand individual Newsfeed postings based on criteria, resulting in views that e.g., show only comments posted by course admins, only new postings, or only top level postings. Especially the 'new comments' filter was highly effective, showing all contributions posted since this filter was last activated.

We added an opportunity to up- and down-vote individual postings, inspired by social media sites like reddit.com. Additionally, we added an element that allowed comments to be 'bookmarked', with an aggregation of all bookmarked comments in a separate view. These changes were effective for all instances of commenting in Aurora, not only in the Newsfeed.

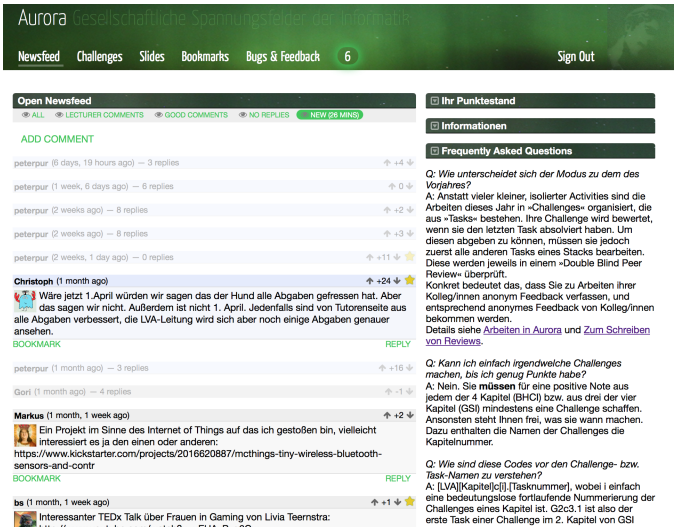


Fig. 9. The Dashboard in the new version of Aurora. The left column is allocated to the Newsfeed, the right column offers the other dashboard widgets such as the points overview and general information widgets (collapsed) and the FAQ widget. The notifications widgets was granted a privileged spot in the menu bar of Aurora, here indicating 6 unseen notifications.

One of the minor changes that penetrated the whole system was the optimization of the overall page dimensions to a width of 960px to make Aurora better suited for mobile screens.

B. Slides

The Slides module was effectively unchanged (Fig. 10). Due to technical problems we were unable to reactivate some of the functions of the Slides module in time for the course. Specifically, the Livecasting component did not work, so that comments could only be added after a lecture was finished.

As a consequence, the amount of comments posted in the Slides Studio was much lower than last year.

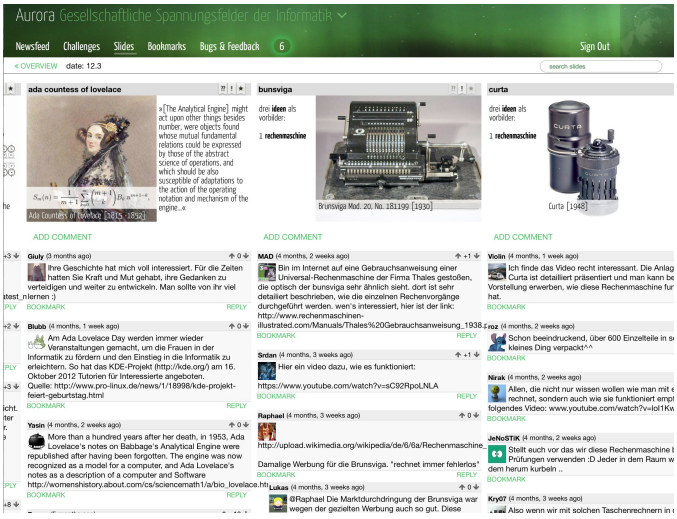


Fig. 10. The Slides Studio in the new version of Aurora.

C. Portfolio

Following the problems with the portfolio approach described above, the principal organisation of the Portfolio module was dropped, and replaced by the 'Challenges' module described below.

D. Discourse

Due to the problems described above, the Discourse module could not be used this year.

VII. Challenges

After abandoning the quite ambitious e-portfolio approach, we introduced a slightly gamified terminology in the following year, with *challenge* as the main metaphor for a complex set of exercises called *tasks*.

In the context of Aurora, a challenge is a chain of exercises called *tasks* where each task (except the first task) requires the completion of the preceding task (Fig. 11). These exercises are usually increasing in their difficulty, and each task is based on the outcome, the skills, or the knowledge gained from the preceding task. Challenges are composed from three to five tasks, with the final task being significantly more work than the preparatory tasks. Typically, a single challenge represents between 10% and 25% of your semester goal.

Students can choose from a catalog of challenges that grows over the semester (Fig. 11 left), following the lecture content. In the end, the catalog comprised of 15 challenges in each course, offering more than 250% of the points necessary to reach the semester goal. The content in each course was divided into four chapters, and students had to choose at least one challenge per chapter to ensure exposure to a balanced set of topics.

As in the previous Portfolio module, we did not set deadlines other than the final end-of-term deadline. Students had to have enough challenges worked through and handed in at the end of the semester in order to fulfill the semester goal. We want to deter students from postponing their work until the end of the semester, so we introduced an organisational constraint, where after handing in a completed challenge you have to wait a number of days until you can hand in another challenge.

The final task represents approximately 50% of the total amount of credits that can be reached in a challenge, and it is evaluated and graded by a staff member or tutor. All preparatory tasks leading up to the final task are submitted into a double blind peer review process, with other course participants as reviewers. Consequently, for every task a student hands in, they have to review three elaborations for the same task handed in by anonymous colleagues. This has the strong pedagogic appeal that you expose students to the work of their peers immediately after they did the same work, leading to not only an exposure to different perspectives on the same material, but also to a guided reflection on their own work.

The double blind peer review process was modelled after the way it is typically organised at conferences; work was assigned randomly to reviewers who were required to answer a couple of questions covering areas such as completeness, correctness, objectivity or originality of the reviewed work. Finally, reviewers were asked to assess the work on a quadrinomial scale ranging from 'Great work' to 'Unacceptable Work', the latter reserved for plagiarism and empty hand-ins.

Participants can work through all preparatory tasks of a challenge without regard for the reviews received. In order to access the final task of a challenge, it is necessary to have at least two positive reviews for each preparatory task.

To maintain the level of quality, tasks as well as reviews were randomly checked by members of the staff. We injected bad and plagiarized work into the review process, and also informed students about it, in order to be able to detect students who systematically refused to invest adequate time into writing their reviews. Also, we implemented an easy way for students to report meaningless reviews they received.

A. Evaluation of Challenges

1) Grading Time: One of the main goals of Challenges was to relieve us of the evaluation overload. Looking back at the semester, we can say that the use of double blind peer reviewing clearly reduced our work load, resulting in a much shorter time-to-evaluation for the students.

If you compare the main pie chart in Fig. 12 with Fig. 5, you will see that the average time for feedback was reduced significantly. We are optimistic that we can reduce it even more, as an organisational mishap in the middle of the semester generated a week of hand-in frenzy, which led to a substantial increase in feedback time; before that week, almost 90% of all hand-ins were evaluated within a week.

Part of that benefit results from a lower number of hand-ins. Instead of up to twenty submissions from each student that had to be graded in the portfolio, we now received six or seven challenges typically handed in by students that subsequently passed. While each challenge consists of three to five separate task elaborations that have to be checked, the reviews attached to all but the final task hand-in helped speed up evaluation and grading significantly.

Additionally, we were now able to provide more substantial feedback to the final task elaboration, as the reduced total number of hand-ins to be evaluated leave more time to compose written feedback. Finally, students no longer complained that it was impossible to engage with some subject matter more in-depth, as challenges provided ample opportunity to delve deeper into any of the offered topics.

2) Peer Reviewing: One of the risks of the introduction of peer reviewing is associated with the fact that students could start any challenge at any time, making it unclear whether they would also receive enough reviews for them to be able to start the final task of the challenge in time. To compensate for this, we had a separate list of all hand-ins that did not receive the necessary two reviews 72 hours after hand-in. Tutors regularly checked that list, providing substitute reviews so that students could advance to the final task of the challenge.

As it turned out, this list was empty most of the time. Only towards the end of the semester did the work of the students

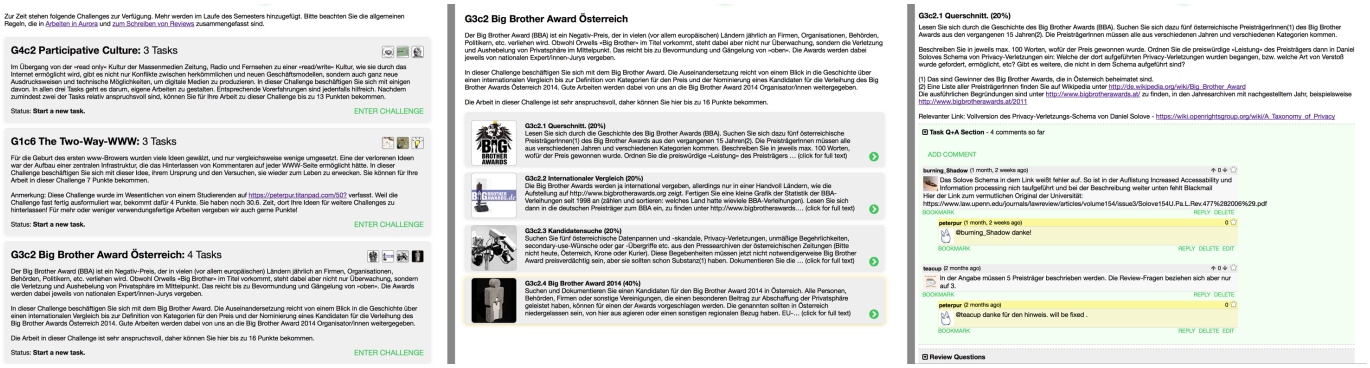


Fig. 11. The three views needed to navigate challenges: [left] Catalog of all challenges (three shown), sorted with latest additions on top; [middle] All tasks in a challenge; [right] A single task view, with the associated Q&A area

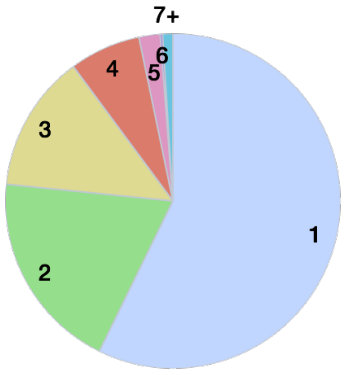


Fig. 12. Time it took to grade an exercise in Challenges, calculated in weeks

diversify so much that they did not receive these minimal two reviews. Of course, the last students to start each task at the end of the semester also had to rely on tutor substitute reviews. In other words, peer reviewing worked very well, and thus most feedback from the students about the peer reviewing mechanism was very positive.

One problem that stood out was the quality of some of the reviews. Some students saw reviewing as an annoying appendage to the core task of working through the challenge, so they wrote predominantly short reviews free of any substantial feedback. We plan to tackle this problem by placing more importance on good reviewing, e.g., by introducing a review reputation value and incorporating this value into the final grade, or by designing an introductory 'meta'-challenge that explains and focuses on the reviewing process.

On the other hand, we received feedback from students who pointed out reviewing as an essential component of the overall experience. The combination of first working through a task, and then reviewing the work of others for that same task was described as very interesting experience, central to what they learned in this course. They also rather liked the reviews they received for their own work, as long as they were substantial enough. Overall, we have the impression that the system leads to a higher involvement in the course, at least for those students who want to get involved.

B. Scope of Challenges

The specific structure of a challenge, being comprised of multiple preparatory tasks escalating into the final task, made it very hard and sometimes impossible to translate a number of the more exotic activities from the Portfolio into the new structure. Especially activities best described as meta-activities like finding new sources, suggesting new content, detecting and correcting mistakes on the slides, or suggesting new activities were hard to incorporate into a challenge. That way, the organisational structure was a step back into more conventional exercise territory. For us, this drawback is more than compensated by the fact that the tasks in the challenge build on each other offer a way to lead students deep into a subject matter, offering guidance and focus.

C. Conclusions

With the structure and organisation of the Challenges module, we believe to have solved a number of our core problems. The delay between hand-in and feedback was down across the board, from the very quick peer reviewing process to the overall evaluation of the whole challenge. As a result, we were able to send students their certificates significantly faster than in any prior year. While quality problems with the double blind peer reviewing were observed as anticipated, we are confident that we can develop concepts to counter those problems.

VIII. Future Plans

As described in Section VII, we consider the Challenges module a huge step in the right direction, and plan to enhance and strengthen it in the suggested ways. We think that we 'cracked' double blind peer reviewing in the context of large university classes.

Slides will probably undergo a major revision for next year. While we consider the general concept to offer a conversational structure following the course content as viable, doubts are mounting whether the slides are in fact the best scaffold for such a structure.

We are already thinking about a better Newsfeed structure, to alleviate the problems observed this year. Due to a much higher conversational 'background noise' in the Newsfeed this

year, many students complained that vital information went under their radar.

The value of having a single channel of communication for a course, without the need to hunt around and look through several modules to find all relevant information and answers cannot be overstated. So far, we have failed in finding a suitable structure, not only but also due to the fact that students never use the offered structures in the way we intend them to. We understand this as a design challenge for the coming years.

Another step in Aurora's development will be to include our discussion component Discourse into the system. The gains from using Discourse to facilitate discussions around the course content were substantial.

Handling more than 500 students in university courses is a rare situation. Often, such a challenge is tackled by introducing distance between teachers and learners, and by relying on examination and tests. This removes autonomy from the learning process, which we see as a central property.

Thus, we tried to go the opposite way, and designed Aurora with the explicit goal to give students as much autonomy as possible in such a setting. In our experience, such a challenge requires explorative approaches, learning not only from evaluation, but also from the design process itself.

ACKNOWLEDGMENT

The authors would like to thank all contributors who have been involved in the development of Aurora over the years: Stephan Bauer, Elisabeth Bauernhofer, Christoph Börner, Daniel Domberger, Michael Emhofer, Andreas Fermitsch, Martin Flucka, Thomas Gradisnik, Peter Holzkorn, Daniel Kececi, Lucia Leitner, Peter Minarik, Wilfried Reinthaler, Gerald Reitschmied, Diane Salter, Reinhard Seiler, Martin Sereinig, Raif Tabucic, Bruno Tunjic, Wolfgang Zalesak. Furthermore, we would like to thank all tutors and students who used Aurora over the years for their invaluable feedback and bug tracking work.

REFERENCES

[1] P. Purgathofer and N. Luckner, "Aurora - Exploring Social Online Learning Tools Through Design," in *Proceedings of The Seventh International Conference on Advances in Computer-Human Interactions (ACHI 2014)*, Barcelona, Spain, 2014, pp. 319–324.

[2] T. O'Reilly, "What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software," *Design*, vol. 65, pp. 17–37, 2007.

[3] S. Downes, "Feature: E-learning 2.0," *Elearn magazine*, vol. 2005, no. 10, p. 1, Oct. 2005, [Accessed Jan. 24, 2014]. [Online]. Available: http://doi.acm.org/10.1145/1104966.1104968

[4] E. Wenger, *Communities of Practice, Learning, Meaning, and Identity*, 1998. [Online]. Available: http://www.stanford.edu/~eckert/PDF/eckert2006.pdf

[5] J. S. Brown and R. P. Adler, "Minds on Fire: Open Education, the Long Tail, and Learning 2.0," *Educause Review*, vol. 43 (1), pp. 16–32, 2008, [Accessed Jan. 24, 2014]. [Online]. Available: http://www.educause.edu/ir/library/pdf/ERM0811.pdf

[6] G. Siemens, "Connectivism: A Learning Theory for the Digital Age," 2004, [Accessed Jan. 24, 2014]. [Online]. Available: http://www.elearnspace.org/Articles/connectivism.htm

[7] M. Irvine, "Survey: 97 Percent Of Children Play Video Games," 2008, [Accessed Jan. 24, 2014]. [Online]. Available: http://www.huffingtonpost.com/2008/09/16/survey-97-percent-of-chil_n_126948.html

[8] J. L. Santos, S. Govaerts, K. Verbert, and E. Duval, "Goal-oriented Visualizations of Activity Tracking: A Case Study with Engineering Students," in *Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge*, ser. LAK '12. New York, NY, USA: ACM, 2012, pp. 143–152, [Accessed Jan. 24, 2014]. [Online]. Available: http://doi.acm.org/10.1145/2330601.2330639

[9] C. Treude and M.-A. Storey, "Awareness 2.0: Staying Aware of Projects, Developers and Tasks Using Dashboards and Feeds," in *Proceedings of the 32Nd ACM/IEEE International Conference on Software Engineering - Volume 1*, ser. ICSE '10. New York, NY, USA: ACM, 2010, pp. 365–374, [Accessed Jan. 24, 2014]. [Online]. Available: http://doi.acm.org/10.1145/1806799.1806854

[10] J. T. Biehl, M. Czerwinski, G. Smith, and G. G. Robertson, "FASTDash: A Visual Dashboard for Fostering Awareness in Software Teams," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '07. New York, NY, USA: ACM, 2007, pp. 1313–1322, [Accessed Jan. 24, 2014]. [Online]. Available: http://doi.acm.org/10.1145/1240624.1240823

[11] M. McKeon, "Harnessing the web information ecosystem with wiki-based visualization dashboards." *IEEE transactions on visualization and computer graphics*, vol. 15, no. 6, pp. 1081–1088, 2009, [Accessed Jan. 24, 2014]. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19834175

[12] S. Govaerts, K. Verbert, J. Klerkx, and E. Duval, "Visualizing activities for self-reflection and awareness," *Advances in Web-Based Learning â ICWL 2010*, vol. 6483, pp. 1–10, 2010, [Accessed Jan. 24, 2014]. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-17407-0_10

[13] E. Duval, "Attention Please!: Learning Analytics for Visualization and Recommendation," in *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, ser. LAK '11. New York, NY, USA: ACM, 2011, pp. 9–17, [Accessed Jan. 24, 2014]. [Online]. Available: http://doi.acm.org/10.1145/2090116.2090118

[14] V. K. Kumar and J. L. Rogers, "Student Response Behaviors in an Instrumented Feedback Environment," *SIGCUE Outlook*, vol. Special, pp. 34–54, Dec. 1978, [Accessed Jan. 24, 2014]. [Online]. Available: http://doi.acm.org/10.1145/1318457.1318461

[15] J. E. Caldwell, "Clickers in the Large Classroom: Current Research and Best-Practice Tips," *CBE-Life Sciences Education*, vol. 6, no. 1, pp. 9–20, 2007, [Accessed Jan. 24, 2014]. [Online]. Available: http://www.lifescied.org/content/6/1/9.abstract

[16] D. Lindquist, T. Denning, M. Kelly, R. Malani, W. G. Griswold, and B. Simon, "Exploring the Potential of Mobile Phones for Active Learning in the Classroom," *SIGCSE Bull.*, vol. 39, no. 1, pp. 384–388, Mar. 2007, [Accessed Jan. 24, 2014]. [Online]. Available: http://doi.acm.org/10.1145/1227504.1227445

[17] S. Teel, D. Schweitzer, and S. Fulton, "Braingame: A Web-based Student Response System," *J. Comput. Sci. Coll.*, vol. 28, no. 2, pp. 40–47, Dec. 2012, [Accessed Jan. 24, 2014]. [Online]. Available: http://dl.acm.org/citation.cfm?id=2382887.2382895

[18] M. Ratto, R. Shapiro, T. M. Truong, and G. W. Griswold, "The Activeclass Project: Experiments in Encouraging Classroom Participation," in *Designing for Change in Networked Learning Environments*, 2003, vol. 2, pp. 477–486.

[19] T. Bergstrom and K. Karahalios, "Social Mirrors as Social Signals: Transforming Audio into Graphics," *Computer Graphics and Applications, IEEE*, vol. 29, no. 5, pp. 22–32, 2009, [Accessed Jan. 24, 2014].

[20] H. Du, M. B. Rosson, and J. M. Carroll, "Augmenting Classroom Participation Through Public Digital Backchannels," in *Proceedings of the 17th ACM International Conference on Supporting Group Work*, ser. GROUP '12. New York, NY, USA: ACM, 2012, pp. 155–164, [Accessed Jan. 24, 2014]. [Online]. Available: http://doi.acm.org/10.1145/2389176.2389201

[21] W. Purgathofer, Peter Reinthaler, "Exploring the Massive Multiplayer E-Learning Concept," *Ed-Media Invited Talk*, pp. 1–9, 2008, [Accessed Jan. 24, 2014]. [Online]. Available: https://igw.tuwien.ac.at/designlehren/exploring_for_edmedia.pdf

[22] L. H. Bryant and J. R. Chittum, "ePortfolio Effectiveness: A(n Ill-Fated) Search for Empirical Support," *International Journal of ePortfolio*, vol. 3, no. 2, pp. 189–198, 2013, [Accessed Jan. 24, 2014]. [Online]. Available: http://www.theijep.com

[23] E. L. Deci and R. M. Ryan, "Motivation, personality, and development within embedded social contexts: an overview of self-determination

theory," in *The oxford handbook of human motivation*, R. M. Ryan, Ed. Oxford, 2012, pp. 85–107.

[24] D. Nandi and M. Hamilton, "How active are students in online discussion forums?" *Proceedings of the Thirteenth Australasian Computing Education Conference*, no. Ace, pp. 125–134, 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=2459952

[25] M. Wilson and C. Fairchild, "Collaborative Learning and the Importance of the Discussion Board," *Journal of Diagnostic Medical Sonography*, vol. 27, no. 1, pp. 45–51, Dec. 2010. [Online]. Available: http://jdm.sagepub.com/cgi/doi/10.1177/8756479310389609

[26] D. R. Comer and J. a. Lenaghan, "Enhancing Discussions in the Asynchronous Online Classroom: The Lack of Face-to-Face Interaction Does Not Lessen the Lesson," *Journal of Management Education*, vol. 37, no. 2, pp. 261–294, Apr. 2012. [Online]. Available: http://jme.sagepub.com/cgi/doi/10.1177/1052562912442384

[27] P. Wright, "The psychology of layout: Consequences of the visual structure of documents," *American Association for Artificial Intelligence Technical Report FS-99-04*, 1999. [Online]. Available: http://www.aaai.org/Papers/Symposia/Fall/1999/FS-99-04/FS99-04-001.pdf

[28] M. C. Dyson, "How physical text layout affects reading from screen," *Behaviour & Information Technology*, vol. 23, no. 6, pp. 377–393, Nov. 2004. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01449290410001715714

[29] S. E. Middlestadt and K. G. Barnhurst, "The influence of layout on the perceived tone of news articles," *Journalism & Mass Communication Quaterly*, vol. 76, no. 2, pp. 264–276, 1999. [Online]. Available: http://jmq.sagepub.com/content/76/2/264.short

[30] D. Popolov, M. Callaghan, and P. Luker, "Conversation Space: Visualising Multi-threaded Conversation," in *Proceedings of the Working Conference on Advanced Visual Interfaces*, ser. AVI '00. New York, NY, USA: ACM, 2000, pp. 246–249. [Online]. Available: http://doi.acm.org/10.1145/345513.345330

[31] D. D. Suthers, "Effects of Alternate Representations of Evidential Relations on Collaborative Learning Discourse," in *Proceedings of the 1999 Conference on Computer Support for Collaborative Learning*, ser. CSCL '99. International Society of the Learning Sciences, 1999. [Online]. Available: http://dl.acm.org/citation.cfm?id=1150240.1150314

[32] S. McCloud, *Reinventing Comics: How Imagination and Technology Are Revolutionizing an Art Form*. William Morrow Paperbacks, 2000.

[33] P. Purgathofer and N. Luckner, "Layout Considered Harmful : On the Influence of Information Architecture on Dialogue," in *Learning and Collaboration Technologies. Designing and Developing Novel Learning Experiences, HCII 2013*, P. Zaphiris and A. Ioannou, Eds. Heraklion, Crete: Springer International Publishing, 2014, pp. 216–225. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-07482-5_21

[34] D. J, *Logic: the theory of inquiry*. H. Holt and Company, New York, 1938.

[35] A. P. Rovai, "A constructivist approach to online college learning," *The Internet and Higher Education*, vol. 7, no. 2, pp. 79 – 93, 2004, [Accessed Jan. 24, 2014]. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1096751604000144

[36] P. Mayring, *Qualitative Inhaltsanalyse, Grundlagen und Techniken*, 8th ed. Weinheim: Beltz, 2003.

A Decentralized Approach for Virtual Infrastructure Management in Cloud Datacenters

Daniela Loreti and Anna Ciampolini
Department of Computer Science and Engineering, Universitá di Bologna
Bologna, Italy
Email: {daniela.loreti, anna.ciampolini}@unibo.it

***Abstract*—The ever growing complexity of modern data centers for cloud computing - mainly due to the increasing number of users and their augmenting requests for resources - is pushing the need for new approaches to cloud infrastructure management. In order to face this new complexity challenge, many organizations have been exploring the possibility of providing the cloud infrastructure with an autonomic behavior, i.e., the ability to take decisions about virtual machine (VM) management across the datacenter's physical nodes without human intervention. While many of these solutions are intrinsically centralized and suffer of scalability and reliability problems, we investigate the possibility to provide the cloud with a decentralized self-organizing behavior. We present a novel migration policy with a twofold goal: saving energy (by putting in sleep mode the underutilized nodes of the datacenter), while keeping the load balanced across the working physical machines. Our migration policy is suitable for a distributed environment, where hosts can exchange status information with each other according to a predefined protocol. We evaluate the performance of the approach by means of an ad hoc built simulator. As we expected, although our distributed implementation cannot perform as good as a centralized management, it can significantly contribute to augment the degree of scalability of a cloud infrastructure.**

Keywords - Distributed Infrastructure Management, Cloud Computing, Self-Organization, Autonomic Computing

I. INTRODUCTION

This article is an extended version of the work [1] presented at ICAS 2014. It reports more detailed explanation of the model, as well as further investigation of the performance of the approach in a simulated senario.

The Cloud Computing paradigm experienced a significant diffusion during last few years thanks to its capability of relieving companies of the burden of managing their IT infrastructures. At the same time, the demand for scalable yet efficient and energy-saving cloud architectures makes the Green Computing area stronger, driven by the pressing need for greater computational power and for restraining economical and environmental expenditures.

The challenge of efficiently managing a collection of physical servers avoiding bottlenecks and power waste is not completely solved by Cloud Computing paradigm, but only partially moved from customers's IT infrastructure to provider's big data centers. Since cloud resources are often managed and offered to customers through a collection of virtual machines (VMs), a lot of efforts concerning the Cloud Computing paradigm are concentrating on finding the best virtual machine (VM) allocation to obtain efficiency without compromising performances.

Since an idle server is demonstrated to consume around 70% of its peak power [2], packing the VMss into the lowest possible number of servers and switching off the idle ones, can lead to a higher rate of power efficiency, but can also cause performance degradation in customers's experience and Service Level Agreements (SLAs) violations.

The operation of turning back on a previously switched off host can be very time-consuming. In modern data centers, aiming to obtain a more reactive system, the underloaded hosts are never completely switched off, but only put into sleep mode. This technique slightly increases the power consumption, but also speeds up the wake up process when other computational power is needed.

On the other hand, allocating VMs in a way that the total cloud load is balanced across different nodes will result in a higher service reliability and less SLAs violations, but forces the cloud provider to maintain all the physical machines switched on and, consequently, causes unbearable power consumption and excessive costs.

In addition, we must take into account that such a system is continuously evolving: demand of application services, computational load and storage may quickly increase or decrease during execution. Due to these contrasting targets, the VM management in a Cloud Computing datacenter is intrinsically very complex and can be hardly solved by a human system administrator. For this reason, it is desirable to provide the infrastructure with the ability to operate and react to dynamic changes without human intervention.

The major part of the efforts in this field relies on centralized solutions, in which a particular server in the cloud infrastructure is in charge of collecting information on the whole set of physical hosts, taking decisions about VMs allocation or migration, and operating to apply these changes on the infrastructure [3], [4]. The advantages of these centralized solutions are well known: a single node with complete knowledge of the infrastructure can take better decisions and apply them through a restricted number of migrations and communications. However, scalability and reliability problems of centralized solutions are known as well. Furthermore, as the number of physical servers and VMs grows, solving the allocation problem and finding the optimal solution can be time expensive, so some other approximation algorithm is typically used to reach a sub-optimal solution in a fair computation time [5]. The adoption of a centralized VM management is even unfeasible in those contexts (like Community Cloud [6], [7]

and Social Cloud Computing [8]), in which both the demand for computational power and the amount of offered resources can change dynamically.

In this work, we investigate the possibility of bringing allocation and migration decisions to a decentralized level allowing the cloud's physical nodes to exchange information about their current VM allocation and self-organize to reach a common reallocation plan. To this purpose, we designed a novel distributed policy, Mobile Worst Fit (MWF), able to both save power (by switching off the underloaded hosts) and keep the load balanced across the remaining nodes as to prevent SLA violations. The policy adopts a decentralized approach: we imagine the datacenter as partitioned into a collection of overlapping neighborhoods, in each of which the local reallocation strategy is applied. Taking advantage from the overlapping, the VM redistribution plan propagates from a local to a global perspective. We analyze the effects of this approach by comparing it with the centralized application of a best fit policy. In particular, we rely on the definition of the Distributed Autonomic Migration (DAM) protocol [9], used by cloud's physical hosts to communicate and get a common decision as regards the reallocation of VMs, according to a predefined global goal (e.g., power-saving, load balancing, etc.).

We tested our approach by means of DAM-Sim, a software that simulates the behavior of different policies applied in a traditional centralized way or through DAM protocol on a decentralized infrastructure.

The article is organized as follows: in Section II, we show the architectural structure of our system, giving an overview of the DAM protocol and focusing on the adopted MWF policy; in Section III, we report the experimental results obtained by means of the DAM-Sim simulator; Section IV focuses on the state-of-the-art of Cloud Computing infrastructure management and Section V describe our conclusions and future works.

II. The Architecture

We present a distributed solution for Cloud Computing infrastructure management, with a special focus on VM migration.

As shown in Fig. 1, each physical node of the system is equipped with a software module composed of three main layers:

- the infrastructure layer, specifying a software representation of the cloud's entities (e.g., hosts, VMs, etc);
- the coordination layer, implementing the DAM protocol, which defines how physical hosts can exchange their status and coordinate their work;
- the policy layer, containing the rules that every node must follow to decide where to possibly move VMs.

The separation between coordination and policy layer allow us to use the same interaction model with different policies. In this way, different goals can be achieved by only changing the adopted policy, while the communication model remains the same. We describe each layer in more detail in the following sections.

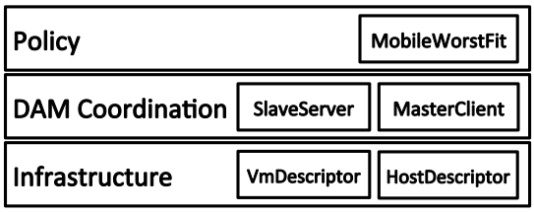


Fig. 1: The three tiers architecture. The separation between layers ensures the possibility to test different policies and protocols with the same infrastructure implementation.

A. Infrastructure Layer

The infrastructure layer defines which information must be collected about each host's status. To this purpose two basic structures are maintained: the *HostDescriptor* and the *VmDescriptor*.

The *HostDescriptor* can be seen as a bin with a given capacity able to host a number of VMs, each one with a specific request for computational resources. We only take into account the amount of computational power in terms of MIPS offered by each host and requested by a VM. An empty *HostDescriptor* represents an idle server that can therefore be put in a *sleep* mode or switched-off to save power.

The *HostDescriptor* contains not only a collection of *VmDescriptor*s really allocated on it (the *current map*), but also a temporary collection (the *future map*) initialized as a copy of the real one and exchanged between hosts according to the defined protocol. During interactions only the temporary copy is updated and, when the system reaches a common reallocation decision, the *future map* is used to apply the migrations.

In a distributed environment, where each node can be aware only of the state of a local neighborhood of nodes, the number of worthless migrations can be very high. Thus, this double-map mechanism is used to limit the number of migrations (as we describe in Section II-B), by performing them only when all the hosts reach a common distributed decision.

Each VM is also equipped with a migration history keeping track of all the hosts where it was previously allocated. For the sake of simplicity, we assume that a VM cannot change its CPU request during the simulation period.

1) The CPU model: The amount U_h of CPU MIPS used by the host h is calculated as follows:

$$U_h = \sum_{vm \in currentVmMap_h} m_{vm} \frac{T_{vm}}{100} \qquad (1)$$

where $currentVmMap_h$ is the set of virtual machines currently allocated on host h; T_{vm} is the total CPU MIPS that a virtual machine vm can request; and m_{vm} is the percentage of this total that is currently used.

Indeed, we consider a simplified model in which the total MIPS executed by the node can be seen as the sum of

MIPS used by each hosted virtual machine. In fact, the model does not take into account the power consumed by the physical machines to realize virtualization and to manage their resources.

B. Coordination Layer

The coordination layer implements the DAM protocol, which defines the sequence of messages that hosts exchange in order to get a common migration decision and realize the defined policy.

The protocol is based on the assumption that the cloud is divided into a predefined fixed collection of overlapping subsets of hosts: we call each subset a *neighborhood*. From an operational point of view, we define a *"knows the neighbor"* relation between the hosts of the datacenter, which allow us to partition the cloud into overlapping neighborhoods of physical machines. As we can see in Fig. 2 the relation is not symmetric.

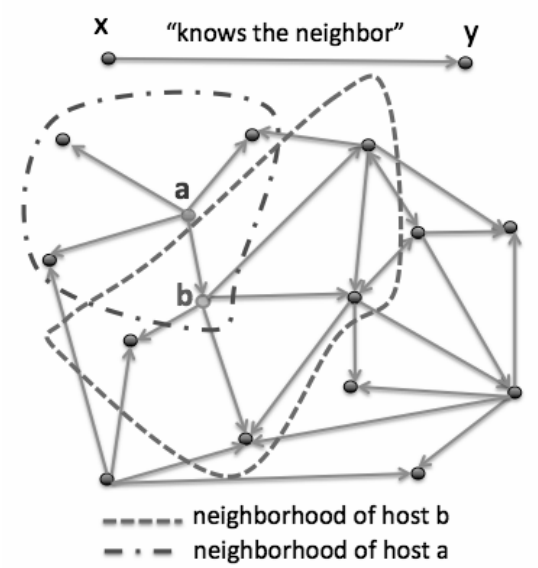


Fig. 2: Example of *"knows the neighbor"* relation applied on a collection of physical nodes. The relation is not symmetric, thus if node *"a knows the neighbor b"*, this does not mean that b is included in the neighborhood of a but, in general, a is not in the neighborhood of b.

We assume that each physical host executes a daemon process called *SlaveServer*, which owns a copy of the node's status stored into an *HostDescriptor* and can send it to other nodes asking for that.

Each node can monitor its computational load and the amount of resources used by the hosted VMs; according to the chosen policy, it can detect either it is in a critical condition or not. A node can, for example, detect to be overloaded, risking to incur in SLA's violations, or underloaded, causing possibile power waste. If one of these critical conditions happens, the node starts another process, the *MasterClient*, to actually make a protocol interaction begin. We call *rising condition* the one that turns on a node's *MasterClient*.

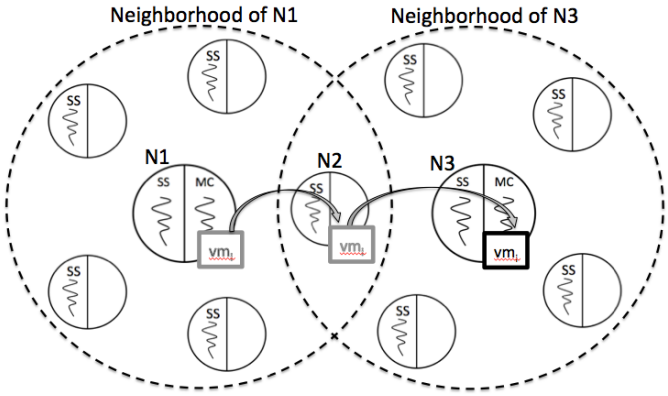


Fig. 3: Schema of two overlapping neighborhoods. The VM descriptor vm_i is exchanged across physical hosts, crossing the neighborhood boundaries, until the nodes agree with a common reallocation plan i.e., a "stable" allocation hypothesis for vm_i is detected.

Since there is a certain rate of overlapping between neighborhoods, the effects of migrations within a neighborhood can cause new rising conditions in adjacent ones.

To better explain the DAM protocol, Fig. 3 shows an example of two overlapping neighborhoods. Each node has a *SlaveServer* (SS in Fig. 3) always running to answer questions from other node's *MasterClient* (MC in Fig. 3), and optionally can also have a *MasterClient* process started to handle a local critical situation. A virtual machine vm allocated to an underloaded node N1 can be moved out of it on N2 and, as a consequence of the execution of the protocol in the adjacent neighborhood of N3, it can be moved again from N2 to N3. It is worth to notice that node N2, as each node of the datacenter, has its own fixed neighborhood, but it starts to interact with it (by means of a *MasterClient*) only if a *rising condition* is observed.

Note that N1's *MasterClient* must have N2 in its neighborhood to interact with it, but the *SlaveServer* of N2 can answer to requests by any *MasterClient* and, if a critical situation is detected (so that N2 *MasterClient* is started) its neighborhood does not necessarily include N1.

As regards this environment, we must remark that the migration policy should be properly implemented in order to prevent never-ending cycles in the migration process.

Algorithms 1 and 2 reports the interaction code executed by the *MasterClient* and the *SlaveServer*, respectively. The *MasterClient* procedure takes as input the list of *SlaveServer* neighbors *ssNeighList* and an integer parameter *maxRound*. The *SlaveServer* procedure takes the *HostDescriptor h* of the node on which it is running. If the *SlaveServer* detects a critical conditions on the host, makes a *MasterClient* process start (lines 1-2 in Alg. 2).

We must ensure that the neighbors's states the *MasterClient* obtains, are consistent from the beginning to the end of the interaction. For this reason, a two-phase protocol is adopted:

1) DAM Phase 1: As we can see in lines 6-10 of Alg.1, the *MasterClient* sends a message to all the *SlaveServers* neighbors *ss* to collect their *HostDescriptors h*. This message also works as a *lock* message: when the *SlaveServer* receives it, locks his state, so that no interactions with other *MasterClients*

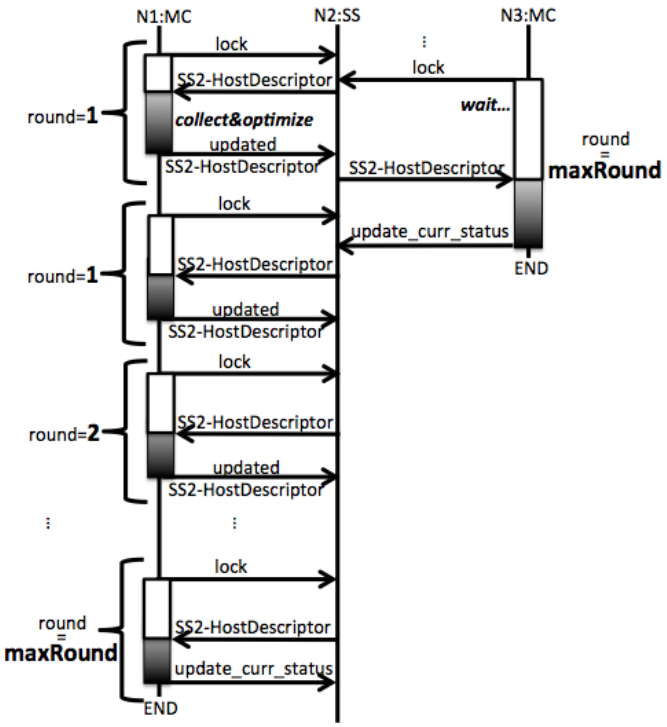


Fig. 4: Example of protocol interaction rounds. Node N2 is shared by nodes N1 and N3. Therefore, their *MasterClients* must coordinate to ensure the consistency of status information.

can take place (lines 6-13 in Alg. 2). If a *MasterClient* sends a request to a locked *SlaveServer*, simply waits for the *SlaveServer* to be *unlocked* and to send its state.

2) DAM Phase 2: The *MasterClient* compares all the received neighbor's *HostDescriptors* with a previous copy he stored (line 12 in Alg. 1). If the *future map* is changed, performs phase 2A, otherwise increments a counter and, when it exceeds a certain maximum, performs phase 2B:

- *Phase 2A:* the *MasterClient* computes a VM reallocation plan for the whole neighborhood, according to the defined policy, and sends back to each *SlaveServer* neighbor the modified *HostDescriptor* (lines 20-23 in Alg. 1). The "optimize(neighHDs)" operation in line 20 of Alg. 1 actually applies the specific chosen policy on the neighborhood's *HostDescriptors* (neighHDs). Indeed, this method is the software connection between the coordination layer and the policy layer.
  As we can see in line 18 of Algorithm 2, the state is accepted passively by the slaves, without negotiation. The migration decisions only change the *future map* of VM allocation. No host switch-on/off or VM migration is performed in this phase. After all new states are sent, the *SlaveServers* are *unlocked* (line 23 in Alg. 2) and the *MasterClient* begins another round of the protocol interaction by restarting phase 1.
- *Phase 2B:* when the number of round with unchanged neighbor's allocation exceeds a defined maximum (line 18 of Alg. 1), the *MasterClient* sends an *update-current-status* request (line 27 of Alg. 1) to all *SlaveServers* and terminates. This last message notifies the *SlaveServers* that information inside the *HostDescriptor* should be applied to the real system state (line 21 of Alg. 2). The *SlaveServer* again executes it passively and unlocks his state.

Algorithm 1 MasterClient DAM protocol code

Input: maxRound, ssNeighList

```
1:  neighHDs = emptylist();
2:  neighHDsPast = emptylist();
3:  round = 0;
4:  while true do
5:    {PHASE 1}
6:    for each ss in ssNeighList do
7:      send(ss, "lock");
8:      (h, ss) = receive();
9:      neighHDs.add(h);
10:   end for
11:   {PHASE 2}
12:   if neighHDs == neighHDsPast then
13:     round++;
14:   else
15:     round = 0;
16:     neighHDsPast = neighHDs;
17:   end if
18:   if round < maxRound then
19:     {PHASE 2A}
20:     optimize(neighHDs);
21:     for each ss in ssNeighList do
22:       send(ss, neighHDs.get(ss));
23:     end for
24:   else
25:     {PHASE 2B}
26:     for each ss in ssNeighList do
27:       send(ss, "update_current_status");
28:     end for
29:     break;
30:   end if
31: end while
```

Alternatives 2A and 2B come from the need for reducing the number of migration physically performed. Looking at example in Fig. 3, if hosts only exchange and update the current collection of VMs, every *MasterClient* can only order a real migration at each round, so that vm_i on N1 would be migrated on N2 at first, and later on N3. Using the temporary *future map* (initially copied from the real one) and performing all the reallocations on this abstract copy, real migration are executed only when the N3's *MasterClient* exceeds a maximum number of rounds and vm_i can directly go from N1 to N3.

The same example is represented in Fig. 4. N2 is shared by the *MasterClient*s of nodes N1 and N3. Two concurrent sessions of the protocol must synchronize in order to maintain the status information consistent. Therefore, node N3 waits until N2 status is updated and released by N1. If no concurrent interactions are taking place in adjacent neighborhoods,

Algorithm 2 SlaveServer DAM protocol code

Input: h

```
1: if checkRisingCondition() then
2:    startMasterClient();
3: end if
4: while true do
5:    {PHASE 1}
6:    (msg, mc) = receive();
7:    if msg! = "lock" then
8:       {protocol error}
9:       break;
10:   else
11:      lock();
12:      send(mc, h);
13:   end if
14:   {PHASE 2}
15:   (item, mc) = receive();
16:   if item! = "update_current_status" then
17:      {PHASE 2A}
18:      h = item;
19:   else
20:      {PHASE 2B}
21:      updateCurrentStatus(h);
22:   end if
23:   unlock();
24:   if checkRisingCondition() then
25:      startMasterClient();
26:   end if
27: end while
```

the *MasterClient* receives an unchanged *HostDescriptor* and increments the value of the *round* counter.

As a result of DAM protocol, the consensus on migration of VMs is not for the entire infrastructure, but is distributed across the neighborhoods. This element must be taken into account while implementing the policy layer.

C. Policy Layer

The Policy Layer is responsible for the decentralized migration decision process. This paper presents MWF, a novel policy aiming to switch off the underloaded hosts to save power, while maintaining the load of the other nodes balanced. MWF exploits two fixed thresholds (FTH_UP and FTH_DOWN) and two dynamic (mobile) thresholds (MTH_UP and MTH_DOWN) used to detect rising conditions. The fixed thresholds identify risky situations: if the host is less loaded than FTH_DOWN an energy waste is detected, while, if the host is more loaded than FTH_UP, SLA violations may occur. The dynamic thresholds (MTH_UP and MTH_DOWN) represents the upper and lower values that cannot be exceeded in order to maintain the neighborhood balanced.

According to the DAM coordination protocol, at each iteration the *MasterClient* collects the VM allocation map of the neighbors and executes a MWF optimization as detailed in

Algorithm 3 MWF policy

Input: h, t, FTH_DOWN, FTH_UP.

```
1: ave = calculateNeighAverage();
2: MTH_DOWN = ave − t;
3: MTH_UP = ave + t;
4: u = h.getLoad();
5: if u < FTH_DOWN or u < MTH_DOWN then
6:    vmList = h.getFutureVmMap();
7: else if u > FTH_UP or u > MTH_UP then
8:    vmList = selectVms();
9: end if
10: if vmList.size ≠ 0 then
11:    migrateAll(vmList);
12: end if
```

Algorithm 4 selectVms() procedure

Input: h, MTH_UP, FTH_UP. **Output:** vmsToMove.

```
1: u = h.getLoad();
2: vmList = h.getFutureVmMap();
3: vmList.sortDecreasingLoad();
4: minU = ∞; bestVm = null;
5: thr = min{FTH_UP, MTH_UP};
6: vmsToMove = emptyList();
7: while u > thr do
8:    for each vm in vmList do
9:       var = vm.getLoad() − u + thr;
10:      if var ≥ 0 then
11:         if var < minU then
12:            minU = var;
13:            bestVm = vm;
14:         end if
15:      else
16:         if minU == ∞ then
17:            bestVm = vm;
18:         end if
19:         break;
20:      end if
21:   end for
22:   u = u − bestVm.getLoad();
23:   vmsToMove.add(bestVm);
24:   vmList.remove(bestVm);
25: end while
```

Alg. 3: the *MasterClient* calculates the average of resource utilization in his neighborhood (*calculateNeighAverage()* in line 1 of Alg. 3) and uses it to compute the two dynamic thresholds (MTH_DOWN and MTH_UP) by adding and subtracting a tolerance interval t (lines 2, 3 of Alg. 3). Then the *MasterClient* checks its *HostDescriptor* h and collects the current computational load u by invoking a specific $getLoad()$ method on the *HostDescriptor* (line 4 of Alg. 3).

The computational load u of the host is compared to fixed and dynamic thresholds: if it is less than the lower thresholds,

Algorithm 5 migrateAll() procedure

Input: vmList, h, offNeighList, underNeighList, otherNeighList.

```
1:  vmList.sortDecreasingLoad();
2:  for each vm in vmList do
3:    vmU = vm.getLoad();
4:    maxAvail = 0; bestHost = null;
5:    for each n in otherNeighList do
6:      if n ∉ vm.getMigrationHistory() then
7:        avail = FTH_UP − n.getLoad() + vmU;
8:        if avail > maxAvail then
9:          maxAvail = avail;
10:         bestHost = n;
11:       end if
12:     end if
13:   end for
14:   if bestHost == null then
15:     minU = ∞
16:     for each n in underNeighList do
17:       if n ∉ vm.getMigrationHistory() then
18:         avail = FTH_UP − n.getLoad() + vmU;
19:         if avail >= 0 and avail < minU then
20:           minU = avail;
21:           bestHost = n;
22:         end if
23:       end if
24:     end for
25:   end if
26:   if bestHost == null and !empty(offNeighList)
      and u > FTH_UP then
27:     bestHost = offNeighList.get(0);
28:   else
29:     migrationMap = null; {all-or-none behavior}
30:     break;s
31:   end if
32:   migrationMap.add(vm, bestHost);
33: end for
34: commitOnFutureMap(migrationMap);
```

the *MasterClient* attempts to put the host in *sleep* mode by migrating all the VMs allocated; otherwise, if the host load exceeds the upper thresholds, only a small number of VMs are selected for migration. As we can see in lines 5, 6 of Alg.3, if the computational load u is less than the fixed (FTH_DOWN) or the dynamic (MTH_DOWN) lower thresholds, all the VMs of the host are collected for migration into an array $vmList$. $h.getFutureVmMap()$ in line 6 is the method to collect the temporary allocation. Indeed in this phase, the policy only works on a copy of the real VM allocation map, because according to DAM protocol, all the migrations will be performed only when the whole datacenter reach a common decision. If the load u is detected to be higher than the fixed (FTH_UP) or dynamic (MTH_UP) upper thresholds, then the $selectVm()$ operation is invoked to pick (from the host h temporary state) only the less loaded VMs whose migration will result in the host load to go back under both MTH_UP and FTH_UP. $selectVm()$ is a modified version of Minimization of Migrations algorithm from Beloglazov et al. [10] and is detailed in Alg. 4. Differently from [10], we select the threshold thr as the minimum between FTH_UP and MTH_UP.

The list of chosen VMs $vmList$ is finally migrated to neighbors by means of a modified worst-fit policy ($migrateAll(vmList)$ in line 11 of Alg. 3). As shown in Alg. 5, the $migrateAll$ procedure takes as input the list of vm to move ($vmList$), the host h where they are currently allocated, the list $offNeighList$ of switched-off hosts in h's neighborhood, the $underNeighList$ of h's neighbors with load level lower than FTH_DOWN, and $otherNeighList$ of all the other neighbors of h. The procedure considers the VMs by decreasing CPU request and, according to the principles of worst-fit algorithm, tries to migrate it to the neighbor n with the highest value of free capacity (lines 2-13 of Alg. 5). If no neighbor in $otherNeighList$ can receive the vm, the $underNeighList$ is considered with a best-fit approach (lines 14-25 of Alg. 5), thus allocating vm on the most loaded host of the list. This ensure that neighbors with CPU utilization near to FTH_DOWN are preferred, while less loaded ones remain unchanged and will be hopefully switched-off by other protocol's interactions. Finally, if neither hosts in $underNeighList$ can receive vm (e.g, because the list is empty), but h is more loaded than FTH_UP, then h is in a risky situation because SLA's violations can occur. Thus a switched-off neighbor is woken up (line 27 of Alg. 5). $migrateAll(vmList)$ operates in a "all-or-none" way, such that the migrations are committed on the future maps (line 34 of Alg. 5) only if it is possible to reallocate all the VMs in the list (i.e., without making other hosts to exceed FTH_UP), otherwise no action is performed (line 29 of Alg. 5).

As shown in Fig. 5, suppose that a protocol execution by the *MasterClient* of h_b decides to migrate a virtual machine vm_i currently allocated on h_c to h_b. When the *SlaveServer* of h_b is unlocked, the policy execution on h_a's *MasterClient* can decide to put vm_i into h_a. Now if h_c has a *MasterClient* running, and decides to migrate vm_i back to h_c, then h_c can take the same decision as before and a loop in vm_i migration starts. If this happens, the distributed system will never converge to a common decision. In order to face this problem, the MWF policy exploits the migration history inside each *VmDescriptor* to avoid loops in reallocation: a VM can be migrated only on a host that it never visited before. Once the distributed autonomic infrastructure reached a common decision, the migration history of each VM is deleted.

III. Experimental results

To understand the effectiveness of the proposed model, we tested it on DAM-Sim [11], a Java simulator able to apply a specific policy on a collection of neighborhoods through DAM protocol and to compare its performance with a centralized policy implementation.

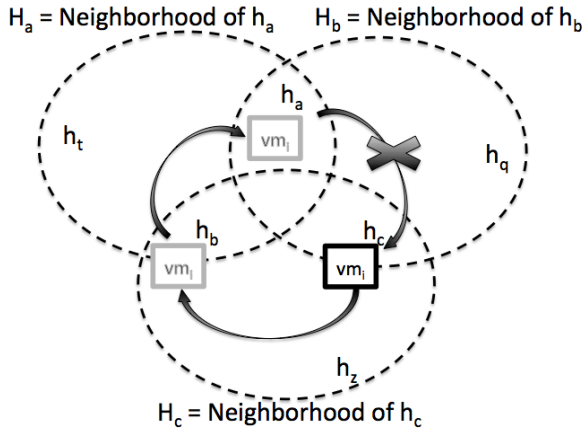


Fig. 5: Example of three overlapping neighborhoods.

We tested our approach on a set of 100 physical nodes hosting around 3000 VMs (i.e., an average value of 30 VMs on each host), repeating every experiment with an increasing average load on each physical server.

According to the tuning tests of Minimization of Migration algorithm [5], the FTH_DOWN and FTH_UP thresholds have been fixed at 25% and 95%, respectively, while we initially set the tolerance interval t for load balancing at 8%.

We start from the worst situation for power-saving purposes, i.e., all the servers are switched on and have the same computational load within the fixed thresholds. To make the DAM protocol start we need some lack of balance in the datacenter, so we forced 20 hosts to be more loaded and 20 hosts to be less loaded than the datacenter average value. These hosts are randomly chosen in every experiment.

In Fig. 6, we compare the MWF performance with nN=5 and 10 nodes in each neighborhood, with the application of a centralized best fit policy (BF-GLO in Fig. 6a) [9]. We also show the performance of BF: a best fit policy applied in a distributed way by means of DAM protocol. BF exploits the two-phase lock protocol, therefore, each time a server detects to be underloaded or overloaded, it start reconsidering the current VM allocation for the whole neighborhood. Details about BF implementation can be found in [9].

Figs. 6a and 6d show the number of servers switched on at the end of the MWF and BF executions. As we expected, the DAM protocol cannot perform better than a global algorithm. Indeed, the global best fit policy can always switch off a higher rate of servers resulting in the lower trend. Furthermore, as regards the power saving objective, we can see that BF perform better than MWF for all the selected neighborhood dimensions. This comes from the different objectives of the two policies: MWF tries to switch-off the initially underloaded servers to save power, while keeping the load of the working servers balanced; BF brings into question all the neighborhood allocation at each *MasterClient* interaction, considering only power-saving objectives.

Figs. 6b and 6e show the number of migrations executed. Since the number of VMs can vary a bit from a scenario to another and the number of switched off servers influences the result, in the graph we show the following rate:

$$nMig\frac{onServers}{nVM} \tag{2}$$

where $nMig$ is the number of migrations performed, $onServers$ is the number of working servers at the end of the simulation and nVM is the number of VMs in the initial scenario.

Since no information about the current allocation of a VM is taken into account during the policy computation in a global environment, the number of migrations can be very high. Indeed is high the resulting trend of migration for the global policy, while DAM always outperforms it. In particular, MWF performs better than BF for every selected neighborhood dimension. Nevertheless, for high values of computational load the performance of MWF in terms of number of switched off server are comparable to those of the global best fit policy, while the migration rate is significantly lower.

Figs. 6c and 6f show the number of messages exchanged between hosts during the computation. As we expected, it significantly increases as the number of servers in each neighborhood grows. Even if the number of messages for low values of neighborhood dimension is comparable to the one of the global solution, when it grows, the number of messages exchanged significantly increases.

In Fig. 7, we can see the distribution of number of servers along load intervals. In the initial scenario (INI in Fig. 7) all the servers have 50% load except for 20 underloaded and 20 overloaded nodes.

The application of a global best fit switches-off a large number of servers to save power, but packs too much VMs on the remaining hosts. This results in the red distribution in Fig. 7, where almost all the switched-on servers are at 95% of utilization, creating an high risk of SLAs violations. The best fit (BF) algorithm applied by means of DAM protocol suffers of the same problem: a large number of servers is switched-off, but a part is forced to have 95% load. MWF is more effective from the load balancing perspective: it can switch-off less servers than BF, but is able to decrease the load of the overloaded nodes leaving all the working servers balanced.

As we expected, Fig. 7 reveals that the median of the MWF distribution is augmented respect to the initial configuration. This is due to the fact that a certain number of servers is switched-off, thus the global load of the remaining servers results increased.

In order to provide a clearer idea of the efficacy of our approach, we separately tested MWF performances in terms of energy efficiency and load balancing. To this purpose, we built three different scenarios.

A. Scenario 1: load balancing test

Considering MWF from the load balancing perspective only, we created a collection of 50 initially unbalanced scenarios satisfying the constraint:

$$U_{TOT} > FTH\_UP(N-1) \tag{3}$$

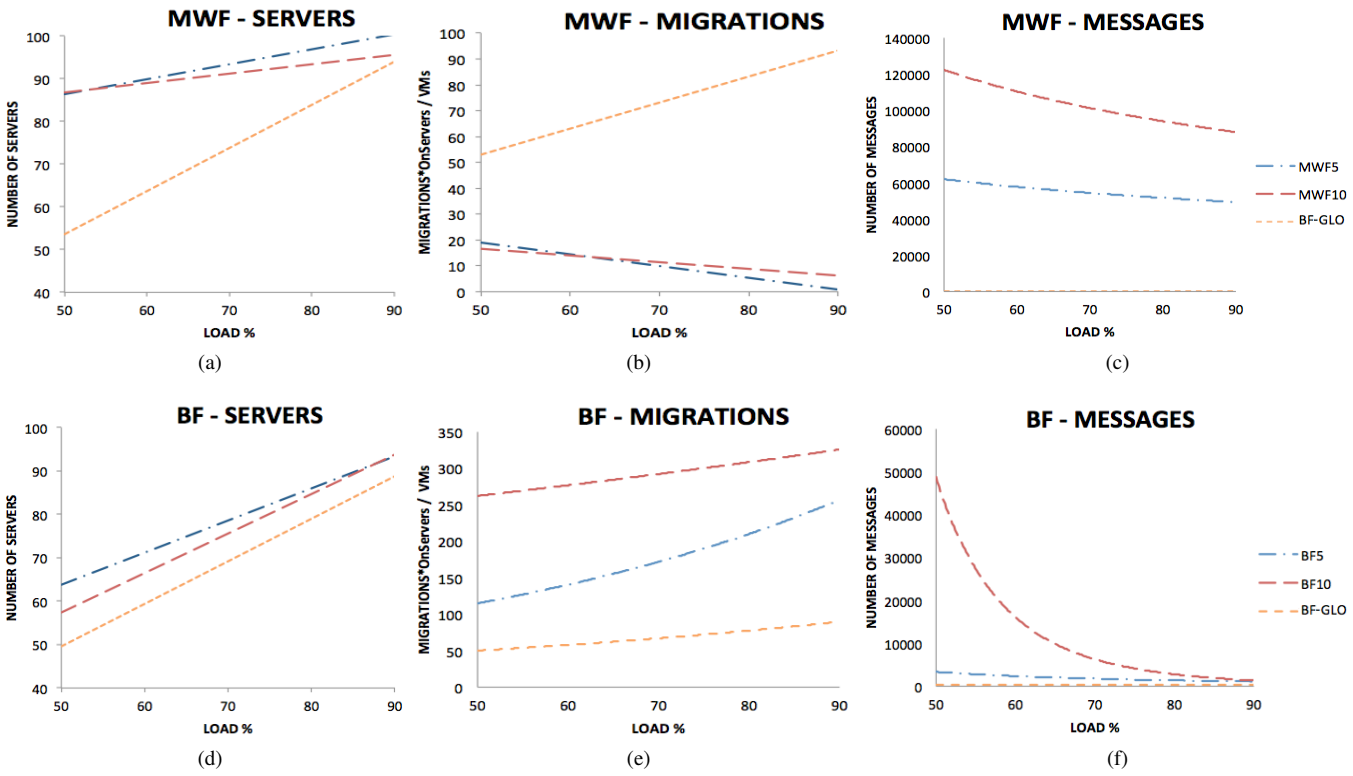


Fig. 6: MWF end BF performance comparison.

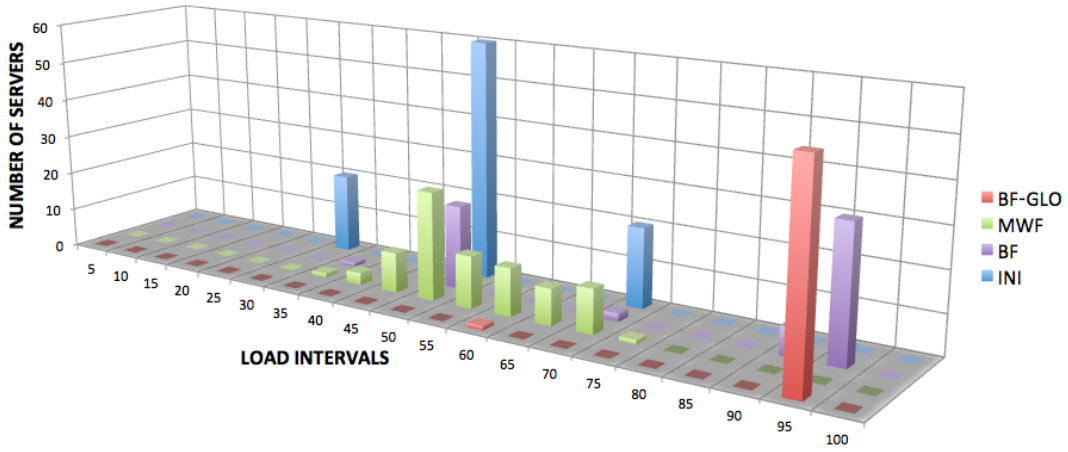


Fig. 7: Distribution of servers on load intervals.

where U_{TOT} is the total CPU-utilization of the datacenter and N is the number of simulated servers. In this way, we can ensure that no server switch-off is possible, and we can test the MWF load balancing performance only.

By defining $U_{AVG\_N} = U_{TOT}/N$ the average load over N servers, the relation 3 can be rewritten as follows:

$$U_{AVG\_N} > FTH\_UP\frac{N-1}{N} \quad (4)$$

and the initial scenarios can be built such that each server h has a CPU-utilization U_h uniformly distributed in the interval:

$$U_h \in [U_{AVG\_N} - q, U_{AVG\_N} + q] \quad (5)$$

where q expresses the degree of imbalance in the initial scenario. We tested the MWF performance with $FTH\_UP = 90\%$, $q = 10\%$ and averaged the results over 50 simulations.

In each scenario the topology of the neighborhoods is generated randomly.

Fig. 8 shows the distribution of servers over load intervals. In the initial scenario (INI) all the servers are on average loaded around the value of FTH_UP. We show the distribution after a global worst-fit optimization (WF-GLO in Fig. 8) and the application of MWF by means of DAM protocol with 10 as neighborhood size.

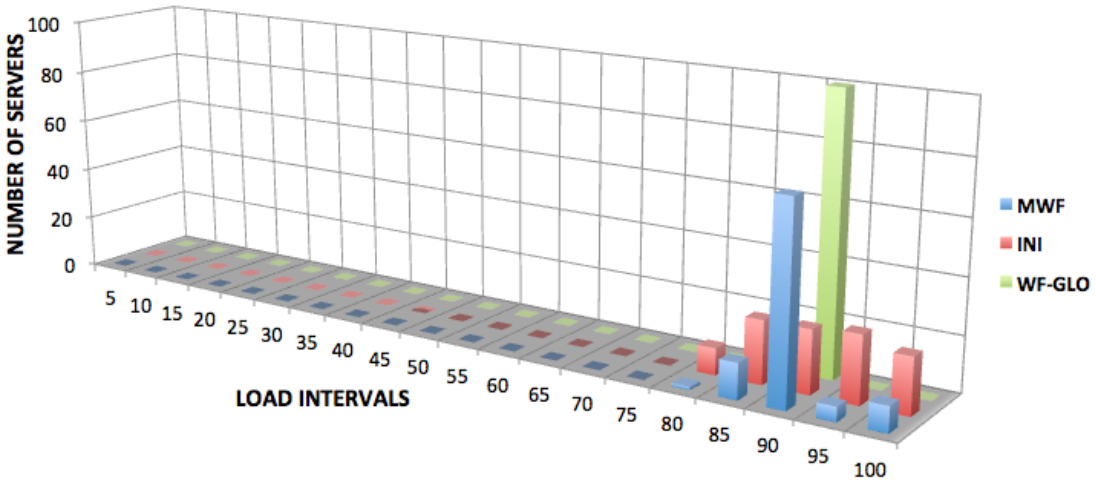


Fig. 8: Distribution of servers on load intervals.

MWF shows good performance from the load balancing perspective even if, as we expected, relying on a global knowledge of status of each server, the centralized application of a worst fit policy clearly outperforms the distributed approach.

B. Scenario 2: power saving test

In order to mainly test the energy saving performance of MWF, we create a collection of scenarios satisfying this constraint:

$$U_{TOT} = FTH\_UP \cdot M, s.t. M < N \tag{6}$$

where M is the number of servers that remains switched-on at the end of the optimization. Relation (6) can be rewritten as follows:

$$U_{AVG\_N} = FTH\_UP \cdot \frac{M}{N} \tag{7}$$

Therefore, given a certain U_{AVG_N} we can calculate the minimum number M_{opt} of servers that can execute the data-center's workload:

$$M_{opt} = \frac{U_{AVG\_N} \cdot M}{FTH\_UP} \tag{8}$$

We create a collection of scenarios with increasing values of U_{AVG_N}, having the load U_h of each server again uniformly distributed in the interval (5) and $q = 20\%$, and we use M_{opt} to evaluate the performance of MWF.

Figs. 9a, 9b and 9c show the number m of working servers at the end of different MWF distributed executions. These values are compared to the minimum possible number M_{opt} of running servers in each scenario.

Each point in the graphs of Fig. 9 represents an initial scenario with different value for U_{AVG_N}.

We repeated the experiment with three different values of the rate $^q/_t$. Fig. 9a shows the energy saving performance with $q = 15\%$ of imbalance in the initial scenario and $t = 5\%$ as MWF tolerance interval. The number of switched-off servers is far from the optimum value (expressed by the blu line) for every generated scenario, while decreasing the ratio $^q/_t$ to $^{20}/_5$ and $^{20}/_3$ (as reported by Figs. 9b and 9c) the performance of MWF significantly increases.

For low values of U_{AVG_N} the algorithm seems to perform significantly better for every value of the rate $^q/_t$, This effect is due to the FTH_DOWN, which is fixed at 25% in every scenario and can therefore contribute to make some $MasterClients$ start if the hosts are detected to be underloaded ($U_h < FTH\_DOWN$).

At the moment, the simulator is not able to give trustworthy results about execution time for distributed environments, because the CPU executing the simulator code can only sequentialize intrinsically concurrent processes of the protocol. For this reason, no test about execution time is reported.

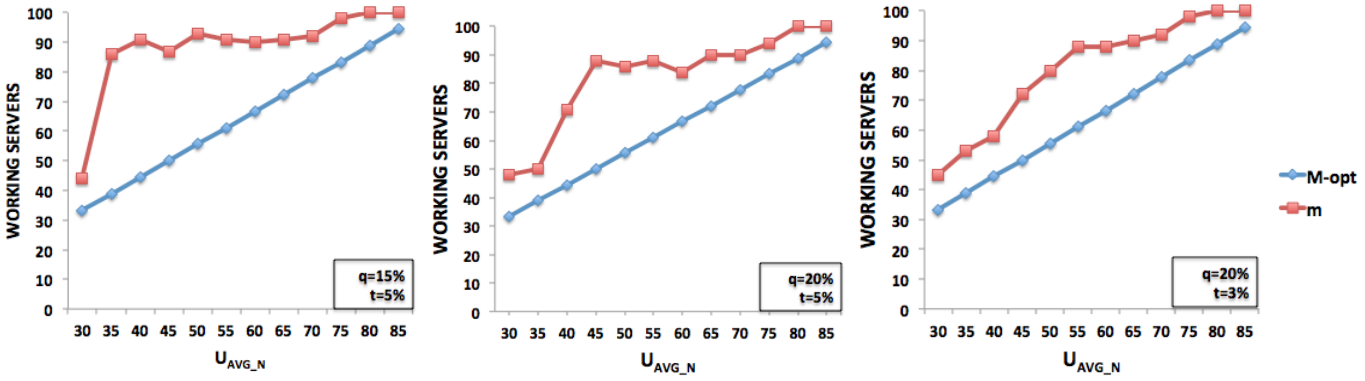
C. Scenario 3: scalability test

In order to test the scalability of the distributed approach, we analyzed MWF behavior while increasing the number of simulated servers and VMs up to 2000 and 60000, respectively. Fig. 10a shows the number of migrations stated by MWF and compare it with that of WF-GLO policy.

Since the number of VMs increases with the number of hosts, we actually compare the ratio between migrations and number of VMs in the scenario.

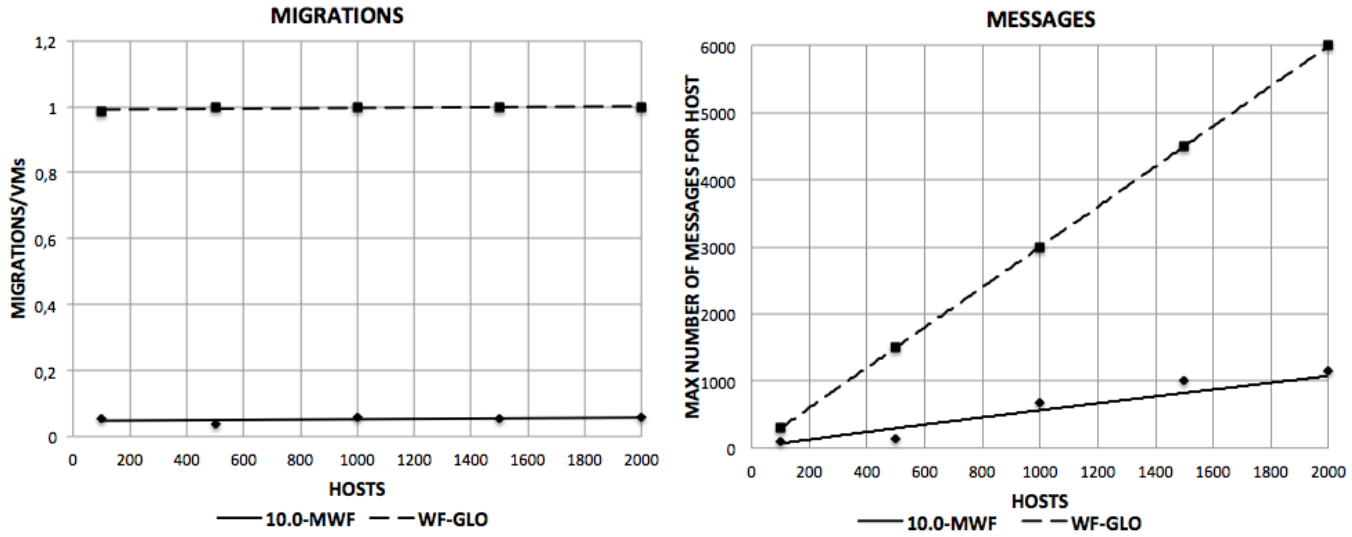
WF-GLO policies does not take into account the current allocation while performing the optimization, therefore, it results in a very high number of migrations, near to the total of VMs. Conversely, MWF distributed policy only operates on underloaded or overloaded nodes (with CPU utilization lower than FTH_DOWN or higher than FTH_UP, respectively) or on those hosts that are unbalanced in respect to the average of their neighborhood (CPU utilization out of the interval $[MTH\_DOWN, MTH\_UP]$). For this reason, as shown in Fig. 10a, the number of resulting migrations is significantly lower for MWF.

In Fig. 10b, we consider the maximum number of messages exchanged by a single host. Since WF-GLO is centralized, the coordinator node must collect the state of all the other nodes before starting the optimization and finally return the



(a) (b) (c)

Fig. 9: MWF power saving performance test. The number m of working servers at the end of different MWF executions is compared to the minimum possible number M_{opt} of running servers in each scenario. The experiments are repeated with different values of the ratio $^{q}/_{t}$.



(a) (b)

Fig. 10: MWF and WF-GLO performance comparison. 10a: Number of migrations performed for increasing number of simulated hosts. We compare the performance of MWF with tolerance interval 10.0 with centralized WF-GLO. 10b: Maximum number of messages exchanged (sent and received) by a single host of the datacenter. MWF significantly outperforms WF-GLO for high values of simulated hosts.

new configuration to each node. Therefore, as shown in Fig. 10b, the number of messages exchanged by the coordinator is always proportional to the number of the nodes it manages. The behavior of MWF policy is again proportional to the number of nodes but the trends is significantly lower. This comes from the fact that, according to MWF, each node of the datacenter always communicate with a predefined number of neighbors (5 in this simulation).

Therefore, considering the maximum number of messages exchanged by a node, for high values of simulated hosts, we can conclude that MWF distributed approach performs better than the centralized WF-GLO algorithm.

IV. Related works

Our work mainly concern low level infrastructural support, in which the management of virtualized resources is always a compromise between system performance and energy-saving. Indeed, in a cloud infrastructure there are usually well-defined SLAs to be compliant to and perhaps the simplest solution is to use all the machines in the cloud. Nevertheless, if all the hosts of the datacenter are switched on, the energy waste increases leading to probably too high costs for the cloud provider.

Around cloud environments, with their contrasting targets of energy-saving versus performance and SLAs compliance, a lot of work was done in order to provide some kind of autonomy from human system administration and reduce complexity. Some of these works involves automatic control theory realizing an intrinsic centralized environment, in which the rate of utilization of each host is sent to a collector node able to determine which physical machines must be switched off or turned on [3], [4], [12]. Some other solutions concern centralized energy-aware optimization algorithms [5], [13]–

[15], in particular extensions of the Bin Packing Problem [16], [17] to solve both VMs allocation and migration problems [10]. These approaches focus on finding the best solution and minimizing the complexity of the algorithm, without concerning the particular implementation, but assuming a solver aware of the whole system state (in terms of load on each physical host and VM allocation). Thus, they particularly lend to a centralized implementation.

Focusing on the SLA compliance perspective only, some recent works have applied the known load balancing techniques (suitable for both distributed systems [18]–[20] and high performance computing [21], [22]) to the Cloud Computing paradigm, particularly as regards the VM allocation problem. A lot of works in this field focus on decentralized solutions [23]–[26] in order to obtain a higher rate of scalability and reliability. As pointed out by Randles et al. [27] comparative study, all these distributed policies again assume that each node can obtain a complete knowledge of the datacenter status. Conversely, our approach does not rely on this strong assumption because each node can work with a local view of the system status, limited by the size of the neighborhood.

Similarly to our approach, in the work by Zhao et al. [28], the decentralized load balancing policy relies not only on the static VM allocation, but also on live migrations in order to run-time dynamically relocate VMs. In [28], as in [27], each node of the datacenter must be able to access a global view of the current allocation scenario.

Furthermore, our work does not focus on the SLA compliance perspective only, but also considers VM consolidation strategies to obtain energy efficiency.

Finally, other approaches involve intelligent, optionally bio-inspired [29], [30], agent-based system, which can give to the datacenter a certain rate of independence from human administration, showing an intelligent self-organizing emergent behavior [31]–[33], and also provide the benefits of a more distributed system structure [34], [35].

As Mastroianni et al. [35] pointed out, building a distributed self-organizing and adaptive infrastructure for VM consolidation can lead to significant scalability performance improvement. As in [35], each physical node of our architecture is able to take decisions on the assignment and migration of VMs exclusively driven by local information. Yet, differently from [35], the assignment procedure of MWF algorithm does not involve all the servers in the datacenter, but only a fixed subset of neighbor nodes.

As in the work by Marzolla et al. [31], which is based on Gossip protocol [36], we adopt a self-organizing approach, where coordination of nodes in small overlapping neighborhoods leads to a global reallocation of VMs, but differently from [31] we created a more elaborate model of communication between physical hosts of the datacenter. In particular, while in [31] each migration decision is taken after a peer-to-peer interaction comparing the states of the only two hosts involved, in our approach the migration decisions are more accurate because they comes from an evaluation of the whole neighborhood state.

V. Conclusions

We presented a decentralized solution for cloud virtual infrastructure management (DAM), in which the hosts of the datacenter are able to self-organize and reach a global VM reallocation plan, according to a given policy. Relying on DAM protocol, we investigated a VM migration approach (MWF) suitable for a distributed management in a cloud datacenter.

Evaluation of MWF policy by means of an ad hoc built software simulator shows good performances for various computational loads in terms of both number of migrations requested and number of switched-off servers. MWF is also able to achieve an appreciable load balancing among the working servers, while still some work remain to do to decrease the number of messages exchanged. Therefore, in the near future, we plan to optimize the DAM protocol in order to reduce the amount of messages in each interaction.

As we expected, the distributed MWF policy cannot outperform a centralized global best-fit policy (especially in terms of number of switched-off hosts and exchanged messages), but further investigations of performance on increasing size datacenters has shown that the decentralized nature of our approach can intrinsically contribute to augment the scalability of the cloud management infrastructure.

In the near future, we will extend DAM-Sim in order to take into account not only computational resources, but also memory and bandwidth requirements. This will allow us to test different and more elaborated reallocation policies. We will introduce variations of VM load requests at simulation time to better mirror real datacenter environments. Furthermore, in this work, we avoid loops in VM migrations by preventing the allocation on nodes that already hosted the same VM before. We plan to relax this restrictive constraint by means of a Most Recently Used queue of hosts.

Further investigation will be necessary to address issues caused by message losses. We will enrich the algorithm with a recovery strategy in order to avoid the risk of physical servers never-ending blocked while they wait for ”unlock” messages.

Finally, we plan to test our implementation on a real cloud infrastructure and compare the time to get a common distributed decision with the centralized implementation of the same reallocation policy. Furthermore, on a real cloud infrastructure we expect to face low level architectural constraints in overlapping neighborhoods definition, which will request deeper investigations.

References

[1] D. Loreti and A. Ciampolini, “Policy for distributed self-organizing infrastructure management in cloud datacenters,” in ICAS 2014, The Tenth International Conference on Autonomic and Autonomous Systems, IARIA, Ed., 2014, pp. 37–43.

[2] X. Fan, W.-D. Weber, and L. A. Barroso, “Power provisioning for a warehouse-sized computer,” in The 34th ACM International Symposium on Computer Architecture. ACM New York, 2007, pp. 13–23.

[3] Jung, “Mistral: Dynamically managing power, performance, and adaptation cost in cloud infrastructures,” in International Conference on Distributed Computing Systems, IEEE, Ed., June 2010, pp. 62–73.

[4] H. C. Lim, S. Babu, and J. S. Chase, "Automated control in cloud computing challeges and opportunities," in ACDC '09, Proceedings of the 1st workshop on Automated control for datacenters and clouds. ACM New York, 2009, pp. 13–18.

[5] A. Beloglazov and R. Buyya, "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers," Concurrency and Computation: Practice and Experience, vol. 24, no. 13, pp. 1397–1420, September 2012.

[6] A. Marinos and G. Briscoe, "Community cloud computing," in First International Conference, CloudCom 2009. Proceedings. Springer Berlin Heidelberg, 2009, pp. 472–484.

[7] C. Giovanoli and S. G. Grivas, "Community clouds a centralized approach," in CLOUD COMPUTING 2013, The Fourth International Conference on Cloud Computing, GRIDs, and Virtualization, IARIA, Ed., 2013, pp. 43–48.

[8] K. Chard, S. Caton, O. Rana, and K. Bubendorfer, "Social cloud: Cloud computing in social networks," in Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on, July 2010.

[9] D. Loreti and A. Ciampolini, "Green-dam: a power-aware self-organizing approach for cloud infrastructure management," Università di Bologna, Tech. Rep., 2013 - http://www.lia.deis.unibo.it/Staff/DanielaLoreti/HomePage_files/Green-DAM.pdf [Accesed 20th November 2014].

[10] A. Beloglazov, J. Abawajy, and R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing," Future Generation Computer Systems, vol. 28, no. 5, pp. 755–768, May 2012.

[11] [Online]. Available: https://bitbucket.org/dloreti/cooperating.cloud.man [Accesed 20th November 2014]

[12] E. Kalyvianaki, "Self-adaptive and self-configured cpu resource provisioning for virtualized servers using kalman filters," in ICAC '09 Proceedings of the 6th international conference on Autonomic computing, ACM, Ed., 2009, pp. 117–126.

[13] R. Jansen, "Energy efficient virtual machine allocation in the cloud," in Green Computing Conference and Workshops (IGCC), 2011 International. IEEE, July 2011, pp. 1–8.

[14] A. J. Younge, "Efficient resource management for cloud computing environments," in Green Computing Conference, 2010 International. IEEE, August 2010, pp. 357–364.

[15] J. C. adn Weidong Liu and J. Song, "Network performance-aware virtual machine migration in data centers," in CLOUD COMPUTING 2012 : The Third International Conference on Cloud Computing, GRIDs, and Virtualization, IARIA, Ed., 2012, pp. 65–71.

[16] J. Levine and F. Ducatelle, "Ant colony optimisation and local search for bin packing and cutting stock problems," Journal of the Operational Research Society, pp. 1–16, 2003.

[17] S. Zaman and D. Grosu, "Combinatorial auction-based allocation of virtual machine instances in clouds," in 2010 IEEE Second International Conference on Cloud Computing Technology and Science (CloudCom), IEEE, Ed., December 2010, pp. 127–134.

[18] A. Piotrowski and S. Dandamudi, "A comparative study of load sharing on networks of workstations," in Int. Conf. on Parallel and Distributed Computing Systems, 1997, pp. 458–465.

[19] A. N. Tantawi and D. Towsley, "Optimal static load balancing in distributed computer systems," Journal of the ACM, vol. 32, no. 2, pp. 445–465, April 1985.

[20] T. Chou and J. Abraham, "Load balancing in distributed systems," IEEE Transactions on Software Engineering, vol. 8, pp. 401–412, July 1982.

[21] G. Aggarwal, R. Motwani, and A. Zhu, "The load rebalancing problem," Journal of Algorithms, vol. 60, no. 1, pp. 42–59, July 2006.

[22] B. J. Overeinder, P. M. A. Sloot, R. N. Heederik, and L. O. Hertzberger, "A dynamic load balancing system for parallel cluster computing," in Future Generation Computer Systems, vol. 12, 1996, pp. 101–115.

[23] F. Saffre, R. Tateson, J. Halloy, M. Shackleton, and J. L. Deneubourg, "Aggregation dynamics in overlay networks and their implications for self-organized distributed applications," The Computer Journal, vol. 52, no. 4, pp. 397–412, 2009.

[24] O. A. Rahmeh, P. Johnson, and A. Taleb-bendiab, "A dynamic biased random sampling scheme for scalable and reliable grid networks," INFOCOMP - Journal of Computer Science, vol. 7, no. 4, pp. 1–10, December 2008.

[25] S. Nakrani, C. Tovey, S. Nakrani, and C. Tovey, "On honey bees and dynamic server allocation in internet hosting centers," Adaptive Behavior, vol. 12, pp. 223–240, 2004.

[26] T. Suzuki, T. Iijima, I. Shimokawa, T. Tarui, T. Baba, Y. Kasugai, and A. Takase, "A large-scale power-saving cloud system with a distributed-management scheme," in International Journal on Advances in Intelligent Systems, vol. 7. IARIA, 2014, pp. 326–336.

[27] M. Randles, D. Lamb, and A. Taleb-Bendiab, "A comparative study into distributed load balancing algorithms for cloud computing," in Advanced Information Networking and Applications Workshops (WAINA), 2010 IEEE 24th International Conference on, April 2010, pp. 551–556.

[28] Y. Zhao and W. Huang, "Adaptive distributed load balancing algorithm based on live migration of virtual machines in cloud," in INC, IMS and IDC, 2009. NCM '09. Fifth International Joint Conference on. IEEE, 2009, pp. 170–175.

[29] R. Giordanelli, C. Mastroianni, and M. Meo, "Bio-inspired p2p systems: The case of multidimensional overlay," ACM Transactions on Autonomous and Adaptive Systems (TAAS), vol. 7, no. 4, p. Article No. 35, December 2012.

[30] S. Balasubramaniam, K. Barrett, W. Donnelly, and S. V. D. Meer, "Bio-inspired policy based management (biopbm) for autonomic communications systems," in 7th IEEE International workshop on Policies for Distributed Systems and Networks, IEEE, Ed., June 2006, pp. 3–12.

[31] M. Marzolla, O. Babaoglu, and F. Panzieri, "Server consolidation in clouds through gossiping," Technical Report UBLCS-2011-01, 2011.

[32] A. Vichos, "Agent-based management of virtual machines for cloud infrastructure," Ph.D. dissertation, School of Informatics, University of Edinburgh, 2011.

[33] A. Esnault, "Energy-aware distributed ant colony based virtual machine consolidation in iaas clouds," Master's thesis, Universitè de Rennes, 2012.

[34] M. Tighe, G. Keller, M. Bauer, and H. Lutfiyya, "A distributed approach to dynamic vm management," in Network and Service Management (CNSM), 2013 9th International Conference on, October 2013, pp. 166–170.

[35] C. Mastroianni, M. Meo, and G. Papuzzo, "Probabilistic consolidation of virtual machines in self-organizing cloud data centers," IEEE Transactions on Cloud Computing, vol. 1, no. 2, pp. 215–228, 2013.

[36] M. Jelasity, A. Montresor, and O. Babaoglu, "Gossip-based aggregation in large dinamic networks," ACM Transaction on Computer Systems, vol. 23, no. 3, pp. 219–252, August 2005.

USEM: A Ubiquitous Smart Energy Management System for Residential Homes

Masood Masoodian*, Elisabeth André†, Michael Kugler†, Florian Reinhart†, Bill Rogers*, and Kevin Schlieper†
*Department of Computer Science
The University of Waikato
Hamilton, New Zealand
Email: {masood, coms0108}@waikato.ac.nz
†Human Centered Multimedia
Augsburg University
Augsburg, Germany
Email: andre@informatik.uni-augsburg.de,
{michael.kugler, florian.reinhart, kevin.schlieper}@student.uni-augsburg.de

***Abstract*—With the ever-increasing worldwide demand for energy, and the limited available energy resources, there is a growing need to reduce our energy consumption whenever possible. Therefore, over the past few decades a range of technologies have been proposed to assist consumers with reducing their energy use. Most of these have focused on decreasing energy consumption in the industry, transport, and services sectors. In more recent years, however, growing attention has been given to energy use in the residential sector, which accounts for nearly 30% of total energy consumption in the developed countries. Here we present one such system, which aims to assist residential users with monitoring their energy usage and provides mechanisms for setting up and controlling their home appliances to conserve energy. We also describe a user study we have conducted to evaluate the effectiveness of this system in supporting its users with a range of tools and visualizations developed for ubiquitous devices such as mobile phones and tablets. The findings of this study have shown the potential benefits of our system, and have identified areas of improvement that need to be addressed in the future.**

Keywords–Building energy management, energy usage monitoring; energy usage visualization; information visualization; mobile user interfaces; user evaluation; persuasive technology.

I. INTRODUCTION

People use energy not only for their basic everyday needs such as cooking, cleaning, warming or cooling their houses, transport, etc. but also to make their lives more comfortable and enjoyable. The worldwide energy consumption is therefore increasing rapidly as the world population grows, and more and more of us demand higher standards of living with better life styles.

However, as the worldwide energy resources are limited, and the available fossil-fuel as our primary source of energy is rapidly dwindling, we need to find alternative means of generating energy through renewable resources, as well as saving energy whenever possible. Since the 1970s, when the first oil crisis happened, many technologies have been proposed and developed to assist consumers with saving energy by, for instance, using it more efficiently. In the past, most technologies have focused mainly on energy consumption by the industry, transport, and service sectors. In more recent years, however, increasing attention has being given to residential energy use.

Strategy adopted by existing systems for management and control of energy consumption in residential homes can be divided into two categories, automated and non-automated (manual). Automated systems use mechanisms for controlling home appliances, so that energy-wasting behavior by users can be partly mitigated (e.g., automatically turning off the heating when windows are left open). Non-automated approach to energy saving in private households, on the other hand, can be broken down into two main parts. First, the residents must identify saving potentials in their household, which requires them to be aware of the energy consumption of individual appliances. Second, once they know how and where energy is actually being used, they need to be assisted and persuaded to change their behavior to reduce their energy usage. Unfortunately, however, there are some challenges in achieving both of these two parts. In relation to the first part, usually people know their overall energy consumption (aggregated for the entire household), because that is what their energy providers bill them for. On the other hand, finding out how much energy each device actually uses or how much of the overall consumption each individual person in a house has consumed (i.e., disaggregated consumption data), is very difficult. As challenging as solving this essentially technical problem may be, overcoming the challenge of changing people's behavior to save energy while still living a comfortable life is even harder.

In this article, we expand on our research work presented at ENERGY 2014 conference [1], which introduced a system called Ubiquitous Smart Energy Management (USEM). This system not only provides an automated solution for reduction of electricity usage, but also caters for the non-automated approach by providing detailed energy consumption information to users, and incorporating various tools to assist and encourage them to change their energy consumption behavior.

We start this article with a review of existing building energy management systems, and related research, with a particular focus on residential homes (Section II). We then describe the design of our Ubiquitous Smart Energy Management system (Section III), its client applications (Section IV), and the current prototype (Section V). This is followed by discussion of a laboratory-based study we have conducted to evaluate the usability of the client applications of USEM (Section VI). Finally, we briefly provide the results of an analysis of the capabilities of USEM against existing guidelines for the design of persuasive technology (Section VII), and draw some conclusions (Section VIII).

II. Energy Management in Residential Homes

There are many existing technologies that target saving energy in residential and commercial buildings. These systems are broadly categorized as Building Energy Management Systems (BEMS). They generally consist of two types of components: the physical hardware used to monitor and control energy consuming devices, and the software components that allow various levels of user interaction with the hardware.

The hardware components can further be categorised into either monitoring, or control hardware. Energy monitoring systems can be single-point or distributed. Single point systems usually provide aggregated consumption data for the entire building, while distributed systems are generally wireless, connect to each energy consuming device at the plug level, and therefore provide disaggregated energy usage data for the individual devices they are connected to. Examples of energy monitoring systems for the residential market include systems such Current Cost [2], the Energy Detective [3] and Wattson [4]. However, these systems only show users how much energy they have consumed in the past, and at best make some general suggestions about how to reduce energy usage in the future. Generally, these types of systems cannot actively control appliances to put energy saving tips into practice.

Control hardware, on the other hand, are mainly used by Building Automation Systems (BAS) to actively control energy consuming devices in a building, usually in combination with a range of sensors that react to their environment (e.g., an air-conditioning system can be automatically switched on/off based on temperature sensor data). Examples of such systems include HomeMatic [5], Gira [6], Intellihome [7], Z-Wave [8], and HomeKit [9].

It should however be noted that the distinction between monitoring and control hardware is gradually diminishing. For instance, the Wattson [4] monitoring system works with Optiplug [10] intelligent sockets to switch on/off devices connected to them depending on the availability of surplus electricity.

Kazmi et al. [11] provide an excellent review and comparison of hardware technology used by existing Building Energy Management Systems. They also discuss how BEMS aim to support users with monitoring their energy consumption, providing real-time feedback to them, and allowing users to automatically or manually control their energy consuming devices. Of particular interest here is the role of feedback in encouraging and supporting residential households in reducing their energy consumption.

Although various studies have shown the importance of feedback in reducing energy consumption [12], [13], [14], [15], [16], there are not many systems to support domestic users with managing and visualizing their energy consumption details, and therefore, easily understandable and persuasive feedback systems are likely to appeal to domestic users [11]. There are however a range of issues that feedback systems need to take into account to make them successful, including those resulting from theoretical frameworks that define consumer behavior (e.g., goal-setting, feedback intervention, etc.) [17], [18], [19].

Based on such theoretical frameworks, Fischer [15] emphasizes that feedback should: be based on accurate consumption data, be provided frequently, involve interaction and choice for households, involve appliance-specific breakdown (i.e., disaggregated), be given over a longer period, involve historical or normative comparisons, be presented in an understandable and appealing way. Similarly, based on their review of intervention studies of residential energy conservation, Abrahamse et al. [12] identify the importance of feedback as an effective strategy for reducing energy consumption, particularly if it is given frequently, is combined with goal-setting, allows comparison, and is supported with rewards.

An important factor to take into account is the presentation of feedback, which is crucial in motivating and altering users' behavior to save energy [11]. In addition to conventional forms of feedback (e.g., in textual format, printed records, etc.), technology can be used to provide feedback in a number of other forms, including graphical visualizations, ambient devices, games and social media. These are briefly reviewed below (also see [20], [21], [22]).

A. Graphical Visualizations

Graphical (statistical) visualizations are widely used for presentation of energy consumption data (for a review see [23]). Vine et al. [24] present a summary of the studies, which have investigated the visualization techniques that are used to present users with information on their energy consumption. They report that the most popular visualization techniques include pie and bar charts. However, the preference for one technique or the other seems to be both user- and context-specific, with different visualization having different effects on influencing users' behavior [25].

B. Ambient Devices

Kim and colleagues [26] outline design requirements for ambient devices to create effective persuasion. In a study they identify ten stages from raising awareness to behavior change and the maintenance of behavioral changes. Based on their findings, they then propose several persuasion methods, including subtle indicators for ambient tracking and visually appealing rewards.

The Energy AWARE Clock [27] is an example of an ambient device that visualizes current and past energy usage of a

household. The three design principles of complexity, visibility and accessibility are used to reduce the complexity of consumption data, make visible "hidden" or "not directly obvious" electricity consuming devices, and have the consumption data easily accessible. A three month user study of nine households showed that the users developed a better awareness of their energy use, and thought about changing their behavior to save energy.

Other ambient devices have been developed to help users save other resources such as water. Examples of these include UpStream [28] and Shower Calendar [29]. Studies of these systems have shown that they lead to reduction in water consumption.

Ham and colleagues [30] conducted a study to see if ambient technology has the capability of persuading people subconsciously. In this study, the participants were asked to rate the energy usage of three devices. The three groups of participants either received supraliminal feedback (presentation of a smiling or sad face for 150 ms), subliminal feedback (presentation of a smiling or sad face for 25 ms) or no feedback at all on their given answers. The feedback was given in the form of smiley faces directly after rating the consumption of a device. The results indicated that both groups with feedback gave more correct answers on average than the group without any feedback. Furthermore, the subliminal feedback group gave comparable answers to the supraliminal feedback group, and they also stated that they had not consciously seen any feedback.

C. Games and Social Media

Several systems have been developed to encourage people to conserve energy and increase their energy use awareness through games and social media. The Power Explorer [31] game tries to help teenagers save energy. This mobile phone game takes into account the changes in energy consumption at home by the players. There are different game elements: habitat, pile and duels. The habitat shows the user's avatar in a virtual climate environment, in which energy usage causes CO_2 clouds to appear, which is bad for the avatar. In the pile view, players can see how they are ranked compared to other players, and in the duels players compete directly against each other. The goal of the duels is to increase the energy awareness about appliances, since players have to adjust their household energy consumption to win. A study of Power Explorer showed that a group of players consumed about 20% less energy than a reference group of non-players.

Other research has focused on integration of home energy feedback into social networks. For instance, Mankoff and colleagues [32] demonstrated integration of energy usage feedback to the MySpace social network to motivate people to conserve energy. Similarly, Foster and colleagues [33] have developed a Facebook application, and have shown in a study that energy consumption can be reduced through social encouragement and competition. Petkov and colleagues [34] expand the idea of social comparison with their social application EnergyWiz, in which users can compare their energy usage with their own history and that of others.

Midden and Ham [35], on the other hand, performed a laboratory-based experiment, in which participants could save energy while using a simulated washing machine. This study showed that social feedback provided by an embodied agent was more effective than just factual feedback about the energy savings made.

These types of social network related systems rely on surveillance and self-monitoring techniques. However, they generally only provide feedback at the household level and not at the individual user's level.

III. Ubiquitous Smart Energy Management

Based on the findings of the reviewed studies, and various recommendations made for designing effective feedback systems for supporting energy conservation in residential homes, as discussed in the previous section, we have iteratively designed and developed our Ubiquitous Smart Energy Management (USEM) system. The aim of this system has been to support the monitoring of energy consumption data, and utilizing this data to allow users to set realistic goals. USEM then attempts to encourage users to achieve these goals by providing them with accurate, real-time, and disaggregated feedback. USEM also aims to provide manual, as well as automatic, control of devices based on users' choices, and to allow them to intervene in operation of the system based on the feedback they receive.

The design of USEM has followed a user-centred design approach. This started by developing a set of personas, and scenarios of use, which allowed us to then identify a set of user requirements for USEM. We used the following personas in our scenarios:

- **Frank** is 38 years old. He works as a clerk in a local car rental company. He has two children and is married to Franny. Frank drives an electric car.
- **Franny** is 35 years old. She used to be a receptionist, but is not working currently so that she could take care of her two children.
- **Max** is a 12-year-old boy. He usually goes to school from 8am to 1pm.
- **Fabienne** is a 5-year-old girl. She usually goes to kindergarten from 9am to 1pm.

Because of the two children, it is important for the family to maintain a comfortable temperature level in the house while the children are at home. They all get up at about the same time and they all need warm water to shower. They have solar panels installed on their roof, but they also rely on energy from the local energy provider in case the panels cannot generate enough energy for an autonomous power supply.

From several scenarios that we developed, we identified the following requirements for USEM, which include a combination of automated and non-automated strategies for energy usage management:

- **Controlling devices:** The system must be capable of controlling devices. For example, turning them on and off, or changing their operating mode.
- **Continuously active tasks:** USEM must support continuously active tasks, such as maintaining a certain room

or water temperature, or recharging an electric car to a certain level.

- **One-time tasks:** One-time tasks run only once when initiated by the user, and have a defined end time. Such tasks include, for examples, washing the laundry, or pre-heating the oven.
- **Task scheduling:** The system has to provide a mechanism for intelligent scheduling and execution of one-time and continuously active tasks based on specific criteria, such as the time-of-use energy prices, energy availability, etc.
- **Measuring consumption:** In order to provide detailed statistics on the energy use of the household, and manage scheduling and execution of tasks, USEM must be able to measure and monitor the energy consumption of all the connected devices.
- **Conditional rules:** The system must support the possibility of defining conditional rules, for instance to perform actions when certain conditions are met. An example of such a rule is turning off the heating when nobody is at home.
- **Remote control:** To control devices remotely, USEM has to provide a mobile interface to access various functions of the system. This would allow users to schedule tasks ahead of time, and then react to any problems, which may occur when tasks are executed by the system.
- **User defined settings:** USEM must provide mechanisms for defining user settings. For examples, keeping a comfortable room temperature level, or saving as much energy as possible.

A. Architecture of USEM

USEM has a modular system architecture [36], consisting of three layers as shown in Figure 1.

- **The Ubiquitous Components Layer** consists of all the individual sensors, actuators and devices connected to the system. Sensors are used to measure environmental factors and energy consumption data, while the actuators are used by the system to control the connected devices.
- **The USEM Middleware Layer** combines the various third-party systems connected to the Ubiquitous Components layer, and exposes a unified platform-independent interface to all the layers above it. It provides basic functionality to control devices and retrieve energy consumption data for specific appliances. Additionally, it can also retrieve information from external sources (e.g., energy prices and weather forecasts).
- **The USEM Client Applications Layer** contains all external applications that communicate with the scheduling component or directly with the USEM middleware. Examples of such applications are provided below.

Figure 2 shows the UML component diagram of the internal structure of the USEM middleware layer. As can be seen, the middleware layer consists of five modules:

- **Hardware Interface Components:** are used to communicate with, and control, physical, as well as virtual,

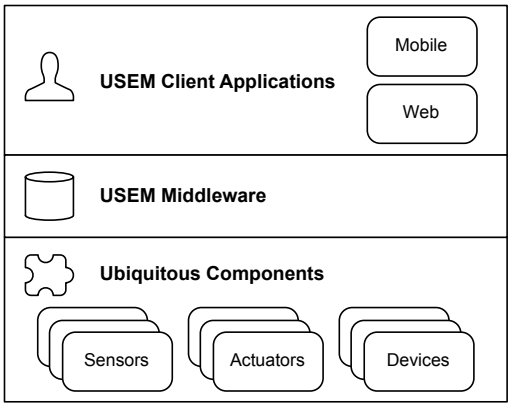


Figure 1: Architecture layers of USEM.

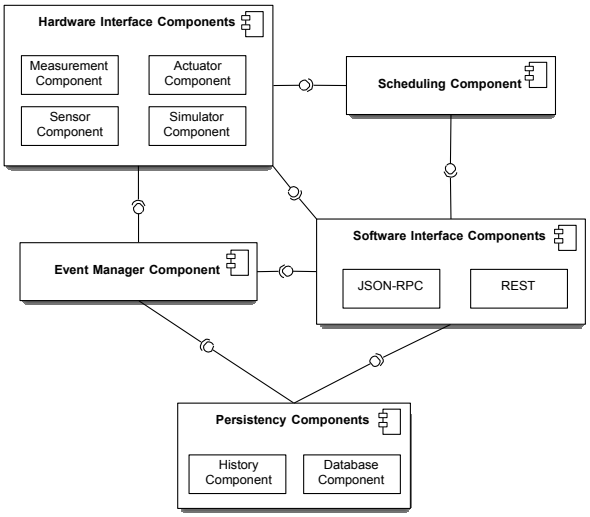


Figure 2: Structure of the middleware components of USEM.

parts of the system connected to the Ubiquitous Components layer. There are four types of such components: energy use measurement devices, actuators, sensors, and simulators.

- **Software Interface Components:** are used by client applications to access the functionality of USEM. We provide two such interfaces. JSON-RPC API [37] can be used to establish a two-way socket connection with the USEM middleware when socket connections are supported. Alternatively, the RESTful API [38] can be used for HTTP communication with the USEM middleware.
- **Persistency Components:** include a history manager, which periodically stores energy consumption data for each device, as well as its operating mode, who has used it, and where it is located in the building. USEM also keeps a database of its devices and hardware components (e.g., which device is plugged into which socket). To better deal with the large amount of data generated,

USEM uses the object-relational mapping framework EclipseLink, which stores the data in a local MySQL database. EclipseLink is an implementation of the Java Persistence API [39] that allows to easily map Java objects to relational database tables.

- **Event Manager Component:** deals with the large number of events generated by different parts of the system. Events are generated, for instance, when the current energy consumption of an appliance changes, when a sensor measures a change in the environment, or when a user manually switches a device on or off, etc. There are also many components that are potentially interested in such events. The history component, for example, must be notified about every change in the consumption value of a device in order to keep track of its consumption history. To deal with all this complexity, we implemented an event handling system based on the observer design pattern [40]. This allows every component interested in receiving a certain event to register for it with the central event manager. Whenever an event occurs, the event manager is notified, so that it can then forward the event and its attached information to any components registered for it. In addition to the original observer design pattern, we not only provide the possibility to register for a single event but also for a whole group of events. For example, a component might be interested in changes that occur in all devices in the living room. With our approach it is possible to register for all devices in the living room at once.
- **Scheduling Component:** is responsible for managing the schedule of all the tasks users have created in USEM. It always tries to find the optimal task execution order. The scheduling component has been implemented using JBoss Drools [41], which allows the creation of rules, as well as workflows and event-processing. USEM uses the JBoss Drools Expert framework to manage the execution of all its rules. The task scheduling and optimization is done by using the JBoss Drools Planner component. Rules and tasks are directly converted into the Drools Rule Language (DRL) format, readable by these components. The scheduler receives commands directly from client applications to create new tasks and to change the calculation parameters. When a scheduled task is due, the scheduling component interacts with the associated appliance through the hardware interface components to set the appropriate operating mode. The scheduler also manages continuously running devices like air-conditioning or water-heating and adheres to the user-defined levels for these devices.

B. The underlying technology

As mentioned earlier, USEM relies on various sensors and actuators connected to its ubiquitous components layer, to manage all the household devices its users would like to control. To do this, we utilize a combination two existing technologies: HomeMatic [5] and Current Cost [2], which are described below.

- **HomeMatic:** is a home automation system designed to control off-the-shelf home appliances and devices. It consists of a central base station (Figure 3, left), which wirelessly communicates with and controls all its power socket adapters (Figure 3, right), to which individual home appliances are connected. The base station also wirelessly communicates with all its sensors. There are different kinds of sensors available: temperature sensors, motion detectors, light sensors, door and window sensors, etc. For controlling off-the-shelf devices, the appliance must be plugged into a HomeMatic adapter socket. There are simple on/off as well as dimmable sockets. Obviously, the limitation of simply cutting the power supply of devices is only suitable for simple devices like lamps, and is insufficient for more complex devices like a washing machine. Such devices must be controlled with a more advanced solution that provides the capability to set different operating modes. The HomeMatic central base station can be controlled either manually using a complex web-based user interface, or automatically through an XML-RPC API [42]. USEM uses this API to communicate with HomeMatic.
- **Current Cost:** was originally designed to monitor and record energy consumption data for an entire household. It can however also be used to monitor individual devices, by attaching a Current Cost jaw device around the power cord of each device (Figure 4, right). The consumption data can then be transmitted wirelessly by each jaw to a Current Cost base station (Figure 4, left). More recently Current Cost has launched a new product (called Individual Appliance Monitor) designed to measure the consumption of individual appliances, which looks similar to a HomeMatic adapter socket. Using either of these tools, energy consumption by each appliance can be measured every six seconds, and stored on the base station. The energy consumption history can be viewed on the display of the base station, or retrieved automatically using a serial data connection via USB. USEM uses this mechanism to access energy use data by devices connected to its ubiquitous components layer.

IV. Design of the Client Applications of USEM

Based on the requirements specified in the previous section we identified a range of functionality to be supported by USEM. To make people aware of their energy usage, USEM would provide detailed statistics about the household's past energy consumption. For example, the overall consumption could be displayed for individual rooms, devices, or occupants of the house. These statistics would allow the users to analyse their consumption history and, thus, identify saving potentials. In some cases USEM might also be able to suggest actions that would lead to a decrease of energy consumption. Furthermore, USEM would support the user in putting theoretical energy saving ideas into practice. For instance, the user could schedule appliances to run when varying energy rates are the cheapest, or USEM could switch off devices when they are not needed (e.g., turning off the printer whenever the PC

Figure 3: HomeMatic base station (left) and socket adaptor for power outlets (right).

Figure 4: Current Cost base station (left) and jaw connected to a power-board (right).

is switched off). By providing an intelligent scheduling for energy consuming tasks USEM would also able to use energy when it is available from off-grid sources (e.g., when the solar panels are generating electricity). Using a combination of these techniques USEM would attempt to ensure that energy usage peaks are avoided, maximum renewable energy is used when available, and overall power usage reduced in an intelligent manner without necessarily reducing comfort levels.

To interact with USEM we designed three different user interfaces: 1) a *web interface* for performing more complex tasks such as managing manual and automatic task scheduling, 2) a *tablet interface* to act as a control unit that could be used from around the house, and 3) a *mobile phone interface* that could be used to interact with USEM while on the go.

For each of these interfaces we identified specific functionalities that they would support. These were then used to design paper prototypes for the three interfaces.

We conducted an expert review of the paper prototypes with four experts, to evaluate the proposed interface designs and functions of USEM. This evaluation resulted in a number of suggestions for improvements, which were used to modify the subsequent versions of the designs used for development of a working prototype.

V. USEM Prototype System

We have developed a functional prototype system based on the modular architecture described earlier. Due to the limitations of currently available home appliances, however, it has been necessary to manually modify some aspects of the interaction between USEM with such appliances. For instance, most existing washing machines cannot be automatically programmed to perform various types of wash cycles, so at this stage we can only turn them on/off, once the user has manually chosen their desired wash cycle. Similarly devices cannot be automatically recognized when connected to USEM, so we print and attach our own barcodes to devices and use these to identify individual devices.

As part of our prototype system, we have also developed three client applications based on the designs discussed in the previous section. These applications allow individual users to login to USEM, not only to interact with the underlying system, but also to allow USEM to monitor energy consumption by individual users, and to provide user-specific controls and information to individual users. Here we present an overview of each of these client applications and their functionalities.

A. Web interface

As mentioned, the web interface of USEM provides higher-level access to the functionality of the system. It allows configuration, scheduling, and visualization of relevant information. Figure 5 shows an example screen of the web interface used for creation of a new task, which will be scheduled and executed automatically by the system. The web interface, along with the more intelligent components of USEM required for the automated energy consumption management strategies (e.g., task scheduling, etc.) have been described more fully elsewhere [43] and will not be discussed further here.

B. Tablet interface

The tablet interface acts as a control and access unit for USEM. Although the current application has been developed for an Apple iPad, it is envisaged that it could also in the future be installed in flat-panel displays incorporated into furniture, picture frames or walls to act as an ambient device interface.

Figure 6 shows the home screen of the tablet interface, which is visible when the device is not being controlled by a user. This allows users to have a constant view of the most important information about their household, which encourages them to monitor their energy use. The home screen is customizable with several widgets to display information such as a list of currently running devices, up-to-date energy prices when available, etc. This screen also shows the energy usage target set by the user, and the current usage level, to motivate the user to keep their usage below their set target. If the target is being threatened, for instance when the user turns

Figure 5: Web interface of USEM, showing a new task being created.

Figure 6: Home screen of the tablet interface of USEM.

on a device, they get a warning from the system giving them the option of turning off another device, which is not being used (if any) or not going ahead with their scheduled activity.

This interface also allows users to view their energy consumption information over the past year, month, or day. Figure 7 illustrates one of the energy usage visualizations. This information can be viewed in several different chart formats, and in various categories, such as for the entire house, different rooms, all users, different users, all devices, different category of devices, etc. This is another important element of the user interface in terms of encouraging energy usage awareness.

In addition, the tablet interface gives energy saving recommendations, based on past and current energy consumption data, to help users reduce their energy use. Figure 8 presents an example energy saving tips screen. On this screen the system

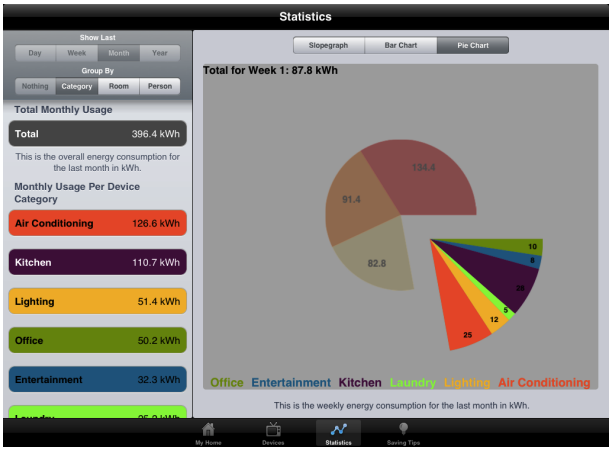


Figure 7: Usage information screen of the tablet interface showing a pie chart.

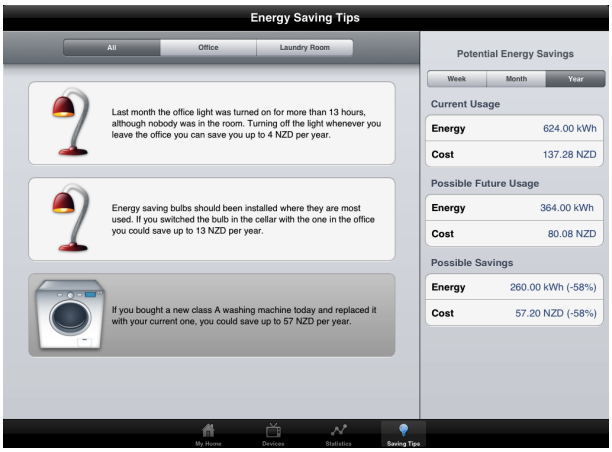


Figure 8: Energy saving recommendations and their consequences if applied.

suggests actions that would decrease the household's energy consumption, as well as calculating the savings that could be made if the advice is followed.

C. Mobile phone interface

The mobile phone interface, developed for Apple iPhone, can be used to retrieve the status of home appliances or to interact with them remotely. It also notifies the user about energy usage events that occur while the user is away (e.g., a scheduled task cannot be undertaken because there is not enough renewable energy available). In such cases, the mobile phone interface provides suggestions (Figure 9) about how the problem could be resolved and gives the user the opportunity to decide what to do (e.g., cancel a scheduled task, or turn off another device).

Of course, the mobile phone interface can also be useful while the user is at home. For instance, it can be used

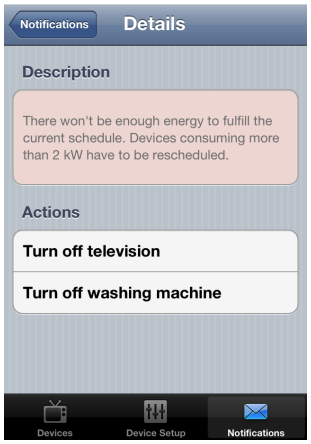


Figure 9: Event notification screen of the mobile phone interface.

Figure 10: USEM devices (left) and new device set-up (right) screens of the mobile phone interface.

to remotely access any of the devices connected to USEM (Figure 10, left). Users can also directly interact with home appliances by scanning the unique barcode that is attached to each device when they are added to USEM. For example, this allows creating a new task for the washing machine right after the user has put the laundry into the machine. The mobile phone application also simplifies the initial set-up procedure for new devices, by mapping the barcode tags attached to power sockets with those of devices connected to them through a simple scanning process (Figure 10, right).

VI. User Study of USEM

We conducted a user study to evaluate the usefulness of USEM and gauge if people would actually use a system like USEM to save energy if they had access to it. In the following sections, we describe this study and discuss some of its main findings.

Figure 11: The set-up used for the user evaluation.

A. Methodology

The study was conducted at a usability lab, where the participants performed a series of tasks using the web, tablet, and mobile phone interfaces. To do the tasks the participants were provided with a laptop, an iPad, an iPhone, and model of a dryer and a computer as two home appliances (each with a barcode attached), as shown in Figure 11. We also attached a barcode to a power socket to make it recognizable by the mobile phone interface of USEM.

Each session started with a tutorial, which included some sample tasks similar to the actual tasks that the participants would perform after the tutorial. Participants were allowed to spend as much time as they needed to complete the tutorial.

The actual study session took about an hour in total, and was divided into three parts covering the use of the three interfaces of USEM. The sessions started with the web interface, as this was the most general part of the system and gave the user a comprehensive overview of USEM (for details of this part of the study and its findings see [43]). This was followed by tasks that users performed using the mobile phone and tablet interfaces.

At the end of each task the participants answered several questions related to the task and the tool they had just used. At the end of the session the participants completed a final questionnaire covering some questions about the users' overall impression of USEM.

B. Study Participants

Twenty participants took part in this study. They were between 20 and 62 years old, with an average age of 35. Five of them were female and 15 male; 11 were students, two researchers, two managers, two office administrators and a housewife. Thirteen of the participants (65%) had previous experience using a multi-touch screen; 11 (55%) owned a smart phone and 4 (20%) owned a tablet device. All of the participants used a computer daily, none had any previous

TABLE I: Demographic of the study participants.

	No. of Participants	Percentage
Gender		
Male	15	75.00 %
Female	5	25.00 %
Occupation		
Student	11	55.00 %
Researcher	2	10.00 %
Other	7	35.00 %
Experience		
Multi-touch screen	13	65.00 %
Daily PC usage	20	100.00 %
Own device		
Smart phone	11	55.00 %
Tablet	4	20.00 %

knowledge of USEM or experience with any other energy management system. Table I shows a summary of the participants' demographic data.

C. Study Tasks

As mentioned earlier, the study participants had to perform specific tasks using each of the different client applications of USEM. We asked the participants to perform the following tasks using the mobile phone interface on the iPhone given to them.

1) *m1: adding a new home appliance to USEM*. In this task the participants were asked to connect a computer given to them to USEM, by physically plugging it into the appropriate power socket, and naming it as "Private PC".
2) *m2: controlling a home appliance remotely*. In this task the participants had to turn off the television remotely.
3) *m3: creating a new scheduled task to let USEM execute it at a later time*. In this task the participants had to create a new task for a given dryer (see Figure 11) by scanning its barcode, and then scheduling the task to be completed by 8am the next morning using the "Delicate" dryer setting.
4) *m4: sending users a demo notification and asking them to react accordingly*. For this task the participants were sent a message (to the iPhone they were using) telling them that there was not enough energy available to perform a dryer task they had previously set up. To resolve this problem they were asked to turn off a device they did not need at that moment (in this case the computer) to make enough energy available for drying the clothes.

The participants then performed the following tasks using the tablet interface on the iPad given to them.

1) *t1: controlling a home appliance remotely*. In this task the participants had to turn on the television. USEM warned the users that they might not achieve their weekly saving goal because some other devices were already turned on. USEM recommended turning off other devices, which the participants then had to do in order to successfully complete the task.
2) *t2: exploring energy consumption statistics using a bar chart visualization*. In this task the participants were asked to use the bar chart to identify the week, in which the household had consumed the highest amount of energy. They also had to specify the amount of energy used during that week.
3) *t3: exploring energy consumption statistics using a pie chart visualization*. This task required the participants to use the pie chart to identify the device category that accounted for the most energy usage during week 4. They also had to specify the amount of energy used by that category.
4) *t4: exploring energy consumption statistics using a slopegraph visualization*. In this task the participants had to find out which device category had the largest increase in energy consumption during week 4 compared to the week before. They also had to specify the amount of this increase. To complete this task, the participants were asked to use the slopegraph. Figure 12 shows an example of type of slopegraph [44], [45] used in this study.

Figure 12: Usage information screen of the tablet interface showing a slopegraph.

D. Task Questionnaires

As mentioned earlier, after the completion of each task the participants were asked to answer a questionnaire. Two questions were common to all the task questionnaires. These were:

1) How easy was it to perform this task?
2) How useful would it be to have this functionality?

The participants answered these questions using a Likert scale of 1-7, with 1 being the least positive and 7 the most positive.

E. Final Questionnaire

Once the participants had completed all the study tasks, they were asked to complete a final questionnaire, which aimed to measure their overall subjective impression of USEM. Table II lists the questions of this questionnaire, along with the 7-point

TABLE II: Questions of the final questionnaire.

No.	Question	Rating	
Q1	How easy were the visualizations on the iPad interface to understand?	1: very difficult	7: very easy
Q2	How likely do you think it would be that you would decrease your energy consumption with the help of the visualizations on the iPad?	1: very unlikely	7: very likely
Q3	Would you want visualizations like on the iPad to be a permanent part of your home?	1: definitely not	7: absolutely
Q4	How easy would it be for you to adapt using USEM for tasks where you do not have to change your daily routine very much? (e.g., create tasks for doing the laundry, instead of just switching the washing machine on manually?)	1: very difficult	7: very easy
Q5	Would you adapt your daily routine in order to use more renewable energy? (e.g., start cooking dinner an hour later?)	1: very unlikely	7: very likely
Q6	How often would you control a device directly or retrieve information about it using the bar code scanner on your mobile phone?	1: very rarely	7: very often
Q7	How often would you use your mobile phone to control your appliances remotely while you are away from home?	1: very rarely	7: very often
Q8	How often would you like to be notified on your mobile phone about what is going on in your household in terms of energy consumption?	1: very rarely	7: very often
Q9	How useful do you find the overall system with regard to efficient energy usage?	1: not useful	7: very useful

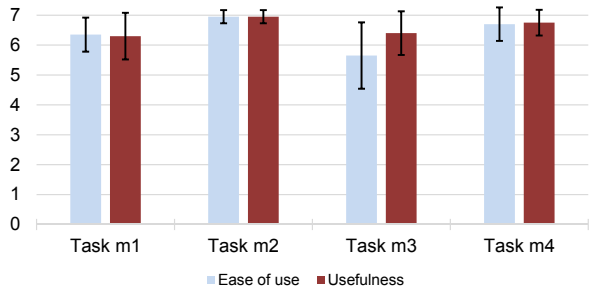


Figure 13: The average ratings given by the participants for the tasks performed using the mobile phone interface.
Note: error bars show the standard deviations.

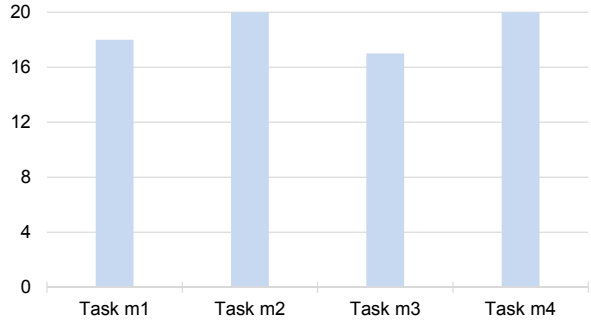


Figure 14: Number of successful completions for the tasks performed using the mobile phone interface.

Likert scales used for each question. The final questionnaire also asked the participants to provide any additional comments or ideas the might have had about USEM.

F. Study Results

1) Results for the mobile phone interface: Figure 13 provides a summary of the average ratings given by the study participants for each of the two questionnaire questions for each of the four tasks performed using the mobile phone interface. As the results show, the participants generally found the tasks easy to perform, and the functionality provided by the interface useful.

Task m3 was considered as being more challenging than the other tasks. However, the functionality needed to perform this task was still rated as being useful. It is also important to note that this task was the most abstract task, which relied on the intelligent scheduling components of USEM.

The ratings given to the difficulty of the tasks is further supported by the number of people who completed each of the tasks successfully. As can be seen in Figure 14, everyone completed tasks m2 and m4 successfully, while 18 people completed Task m1, and 17 people completed Task m3 (this being the most difficult task). Overall the results are very good, considering the fact that the study participants had never used an energy management system previously.

2) Results for the tablet interface: Figure 15 shows a summary of the average ratings given by the study participants for each of the two questionnaire questions for each of the four tasks performed using the tablet interface. Since the difficulty of the tasks steadily increased, the ease-of-use rating for the tasks decreased slightly from Task t1 to Task t4.

However, in general all tasks have been rated as easy to perform with average ratings ranging from 6.30 to 6.75. The ratings given to the usefulness of the functionality provided by the tablet interface for performing each of the tasks showed a trend similar to that observed for the difficulty of the tasks. Once again, overall the participants found the functionality provided very useful.

In terms of the task completion, all the study participants completed all the tasks successfully (see the bars for Part 1 in Figure 16). However, three of the tasks (t2, t3, t4) also had a second part, which asked the participants to report an exact value (in kWh) for energy consumption using one of the three visualizations provided (bar chart, pie chart, slopegraph). As can be seen in Figure 16 (bars for Part 2), all the visualizations were less than perfect in terms of allowing the users to identify the correct energy consumption value, with the pie chart (Task t3) being the worst in accuracy.

3) Results of the final questionnaire: Figure 17 shows the average ratings given by the study participants to each of the questions of the final questionnaire. The results show that the participants had a generally positive view of the various

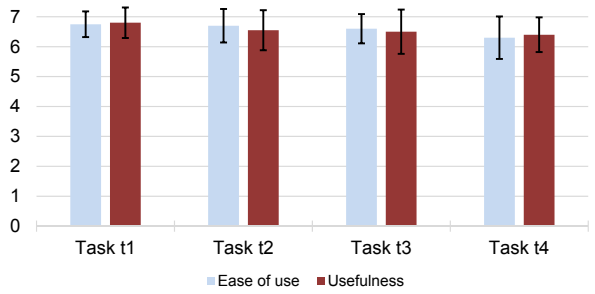


Figure 15: The average ratings given by the participants for the tasks performed using the tablet interface.
Note: error bars show the standard deviations.

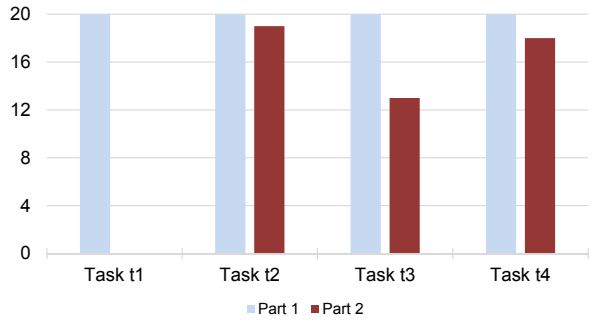


Figure 16: Number of successful completions for the tasks performed using the tablet interface.
Note: Task t1 had only one part.

components of USEM, its client interfaces, and its likely effect in their potential behaviour changes. In particular the participants found the visualizations of the tablet interface easy to use (Question 1), thought these would help them decrease their energy consumption (Question 2), and would like to have them in their homes (Question 3). Although still very positive, the participants were however less committed to using the mobile phone interface to remotely interact with their devices while away from home (Question 7), or receive information about their energy consumption (Question 8). Perhaps the most important finding we can conclude from the final questionnaire is that the participants thought that USEM would be useful in helping them use energy more efficiently (Question 9). It is also important to note that the participants thought they would use USEM to change their daily routines (Question 4) and adapt them in order to use more renewable energy (Question 5).

As mentioned earlier, the final questionnaire also invited the study participants to provide any comments and ideas they might have had about USEM. The following is a summary of some of the main points made by the participants in their comments.

- Many of the participants stated that they would like to be able to determine what events they should be notified about. They feared that they would get annoyed or distracted by notifications if they did not have some control over the notifications sent to their mobile phone.
- Most of the participants especially liked the possibility of controlling all the appliances using a single interface, rather than using a variety of different user interfaces for controlling different home appliances.
- The participants confirmed that they do not know how much energy each of their appliances uses. They rate their energy usage awareness as relatively low. They liked the energy usage visualizations provided by USEM, and thought that these would assist them to better understand the power usage of their appliances.
- Several participants stated that they do not have a good understanding of the kWh measurement unit. Instead they would prefer some kind of visualization, which is easier to understand and does not require any technical knowledge. They also suggested to display dollar amounts, and setting the saving goal in dollars as well.
- One participant commented that he would like recommendations for a saving goal. In this participant's opinion, it is difficult to set a saving goal, since it might be hard to determine a realistic energy consumption limit. So, the system could provide a recommendation for a feasible saving goal based on previous usage data.

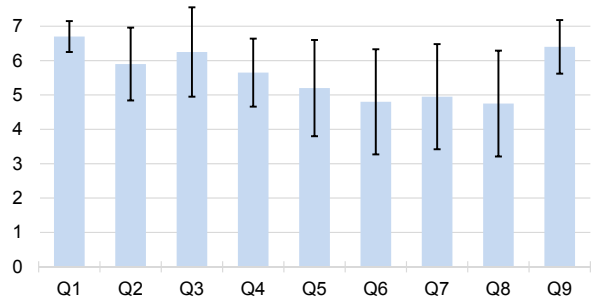


Figure 17: The average ratings given by the participants for the questions of the final questionnaire.
Note: error bars show the standard deviations.

Further to these comments, which are directly related to the functionalities and client application interfaces of USEM, one of the participants raised the issue of security concerns over unauthorised people accessing and controlling their household appliances. Although USEM has not at this stage dealt with the issue of security, in other related research [46] we are investigating security of systems such as USEM.

VII. Persuasive Aspects of USEM

It is important to note that tools and technologies, such as USEM, which aim to assist people with changing their behavior need to be "persuasive" in their approach. The idea of *computers as persuasive technology*, or "captology", was introduced by Fogg [47] to deal with the question of how

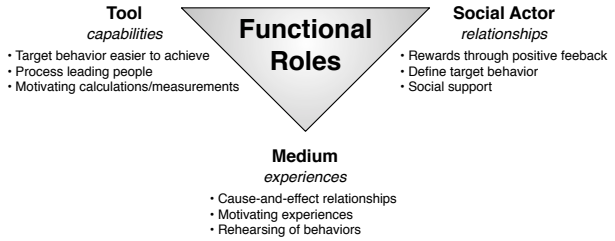


Figure 18: The roles computer technology can play in persuasive context, as defined by Fogg [48].

interactive computer technology could be used to persuade people to change their behavior or attitude. The functional triad, as defined by Fogg [48], is a framework that illustrates three roles computers can play in a persuasive context (Figure 18). These roles are categorized as tool, medium and social actor.

In the context of the work presented in this paper, we are mainly concerned with computers as persuasive tools. Tools increase capabilities by making the desired behavior easier to achieve, by guiding people through processes, or by calculations and measurements that motivate people to reach their goals. There are seven different categories of persuasive technology tools [48], which can be combined together in a single system or application.

1) **Reduction:** People can be persuaded by reducing complexity. A good example of a reduction is the *one-click buy* functionality provided by Amazon [49], which reduces the ordering process to a simple button click.
2) **Tunneling:** This is the process of leading a user step-by-step through a specific procedure. An example of tunneling is the ordering process of online shopping sites. Such a guided process can provide opportunities for persuasion along the way. For instance, an online shopping site can suggest other items of interest to the buyer during the ordering process.
3) **Tailoring:** This approach persuades through customization, by providing only the type of information which is relevant and interesting to the user. An example of this is customized newsletters sent to users offering them products that match their buying profiles.
4) **Suggestion:** This means providing suggestions at the right moment. An example of this is advertisements along a highway, that for instance place an advert for a restaurant near its physical location and not miles away.
5) **Self-Monitoring:** People like to control themselves and check whether they have reached a predetermined goal. An example of this is a heart rate monitor that can be used to monitor the heart rate during exercise.
6) **Surveillance:** People tend to change their behavior when they know that they are being observed. An example of this is messages like "How am I driving?" at the back of some delivery trucks, to ensure that the drivers know people can complain about their bad driving, so they drive more carefully.
7) **Conditioning:** Giving positive, or negative, reinforcement can have a persuasive effect. An example of positive reinforcement is being on the high scorers list in a computer game, which can persuade people to play the game longer to improve their placement on the list.

To measure the success of a system as a persuasive technology clearly requires a long-term study of the use of the system in real-life settings to see if it indeed assists its users with changing their behavior. Although we are yet to conduct such a study of USEM, we have attempted to analyze the ways, in which USEM might be able to play the role of persuasive technology listed above. Below, we provide a summary of this analysis.

1) **Reduction:** USEM reduces the complexity of the large volume of energy usage data, collected for many devices over an extended period, by categorizing it, and allowing the user to view it in a variety of forms.
2) **Tunneling:** USEM provides step-by-step guidance for dealing with the process of adding new appliances to the system, dealing with notifications, managing energy saving targets when they are breached, etc.
3) **Tailoring:** Energy usage information provided is tailored to individual users (i.e., their personal data), energy saving recommendations provided are tailored to each specific USEM installation and are always relevant to the context.
4) **Suggestion:** When USEM warns users about missing their targeted saving goals, it suggests what actions could be taken, for instance by giving a list of devices that could be turned off. Also, when USEM sends notifications to the mobile phone interface when scheduled tasks cannot be undertaken, it provides a list of suggestions that the user can select from.
5) **Self-Monitoring:** By measuring energy usage of each individual (when possible), USEM allows them to monitor their own current performance against targeted saving goals, as well as allowing them to monitor their past usage history in various statistical visualization forms.
6) **Surveillance:** Due to the fine granularity of energy usage data that USEM collects, the user knows (even when living in a house with others) that their consumption behavior is recorded and can be tracked by others when allowed.
7) **Conditioning:** By allowing users to compare their own energy usage behavior to others, as well as their set targets, USEM provides users with positive or negative reinforcements depending on their performance.

VIII. Conclusions

In this paper, we have discussed the design and development of USEM, a system that aims to support the inhabitants of residential homes with the process of monitoring their energy usage, and making energy savings possible without necessarily reducing their comfort levels. USEM allows its users to connect and control their home appliances, as well as analyze and understand energy consumption information by

those appliances using a range of scheduling, notification, and visualization tools.

Our laboratory-based user evaluation of USEM has shown the potential benefits of its client applications, designed for web-browsers, mobile phones, and tablet devices, in providing the necessary means of assisting people with saving energy, as well as encouraging them to monitor and change their energy use behavior.

We have briefly analyzed the capabilities of USEM as a persuasive technology, by examining some the features of its mobile interface components against existing guidelines for the design of persuasive technology. Although this analysis shows that USEM satisfies these guidelines, it is important to conduct a more formal real-life user evaluation of the persuasive capabilities of USEM.

Our study has also identified a range of improvements that could be made to USEM, particularly in improving the type of visualizations it provides to allow the users to effectively access, analyze and compare their energy consumption data. This is an area of interest, which we are investigating further [23].

ACKNOWLEDGEMENTS

The user study presented in this paper was approved by the Ethics Committee of the Faculty of Computing and Mathematical Sciences, The University of Waikato. We would like to gratefully acknowledge the contributions of our study participants.

This research has been supported by the IT4SE project, funded by the German Federal Ministry of Education and Research (Grant number NZL 10/803 IT4SE) under the APRA initiative. More information about the IT4SE project can be found at http://www.it4se.net.

REFERENCES

[1] F. Reinhart, K. Schlieper, M. Kugler, E. André, M. Masoodian, and B. Rogers, "Fostering energy awareness in residential homes using mobile devices." in Proceedings of the 4th International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies, ser. ENERGY 2014. IARIA, 2014, pp. 35–43.

[2] Current Cost Ltd., "Current Cost," http://www.currentcost.com, [retrieved: December, 2014].

[3] Energy Inc., "The Energy Detective (TED)," http://www.theenergydetective.com, [retrieved: December, 2014].

[4] DIY Kyoto, "Wattson," http://www.diykyoto.com, [retrieved: December, 2014].

[5] eQ-3 AG, "HomeMatic," http://www.homematic.com, [retrieved: December, 2014].

[6] Gira Giersiepen GmbH, "Gira," http://www.gira.de, [retrieved: December, 2014].

[7] Intellihome Automatisierungstechnik GmbH, "Intellihome," http://www.intellihome.com, [retrieved: December, 2014].

[8] Sigma Designs Inc., "Z-Wave," http://www.z-wave.com/, [retrieved: December, 2014].

[9] Apple Inc., "HomeKit," https://developer.apple.com/homekit/, [retrieved: December, 2014].

[10] DIY Kyoto, "Optiplug," http://www.diykyoto.com/uk/aboutus/optiplug, [retrieved: December, 2014].

[11] A. H. Kazmi, M. J. O'grady, D. T. Delaney, A. G. Ruzzelli, and G. M. P. O'hare, "A review of wireless-sensor-network-enabled building energy management systems," ACM Transactions on Sensor Networks, vol. 10, no. 4, 2014, pp. 66:1–66:43. [Online]. Available: http://doi.acm.org/10.1145/2532644

[12] W. Abrahamse, L. Steg, C. Vlek, and T. Rothengatter, "A review of intervention studies aimed at household energy conservation," Journal of Environmental Psychology, vol. 25, no. 3, 2005, pp. 273–291. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S027249440500054X

[13] ——, "The effect of tailored information, goal setting, and tailored feedback on household energy use, energy-related behaviors, and behavioral antecedents," Journal of Environmental Psychology, vol. 27, no. 4, 2007, pp. 265–276. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0272494407000540

[14] S. Darby, "The effectiveness of feedback on energy consumption: a review for DEFRA of the literature on metering, billing and direct displays." Environmental Change Institute, University of Oxford, Tech. Rep., 2006. [Online]. Available: http://www.eci.ox.ac.uk/research/energy/electric-metering.php

[15] C. Fischer, "Feedback on household electricity consumption: a tool for saving energy?" Energy Efficiency, vol. 1, no. 1, 2008, pp. 79–104. [Online]. Available: http://dx.doi.org/10.1007/s12053-008-9009-7

[16] G. Fitzpatrick and G. Smith, "Technology-enabled feedback on domestic energy consumption: Articulating a set of design concerns," IEEE Pervasive Computing, vol. 8, no. 1, 2009, pp. 37–44. [Online]. Available: http://dx.doi.org/10.1109/MPRV.2009.17

[17] L. McCalley and C. J. Midden, "Energy conservation through product-integrated feedback: The roles of goal-setting and social orientation," Journal of Economic Psychology, vol. 23, no. 5, 2002, pp. 589–603. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167487002001198

[18] L. McCalley, "From motivation and cognition theories to everyday applications and back again: the case of product-integrated information and feedback," Energy Policy, vol. 34, no. 2, 2006, pp. 129–137, reshaping Markets for the Benefit of Energy Saving Reshaping Markets for the Benefit of Energy Saving. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S030142150400268X

[19] S. Darby, "Literature review for the energy demand research project," Environmental Change Institute, University of Oxford, Tech. Rep., 2010. [Online]. Available: https://www.ofgem.gov.uk/ofgem-publications/59113/sd-ofgem-literature-review-final-081210.pdf

[20] T. G. Holmes, "Eco-visualization: Combining art and technology to reduce energy consumption," in Proceedings of the 6th ACM SIGCHI Conference on Creativity &Amp; Cognition, ser. C&C '07. New York, NY, USA: ACM, 2007, pp. 153–162. [Online]. Available: http://doi.acm.org/10.1145/1254960.1254982

[21] J. Froehlich, L. Findlater, and J. Landay, "The design of eco-feedback technology," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ser. CHI '10. New York, NY, USA: ACM, 2010, pp. 1999–2008. [Online]. Available: http://doi.acm.org/10.1145/1753326.1753629

[22] J. Suppers and M. Apperley, "Developing useful visualizations of domestic energy usage." in Proceedings of the 7th International Symposium on Visual Information Communication and Interaction, ser. VINCI 2014. New York, NY, USA: ACM, 2014, pp. 139–148.

[23] M. Masoodian, B. Endrass, R. Bühling, P. Ermolin, and E. André, "Time-pie visualization: Providing contextual information for energy consumption data," in Proceedings of the 17th International Conference on Information Visualisation, ser. IV '13. IEEE Computer Society, 2013, pp. 102–107.

[24] D. Vine, L. Buys, and P. Morris, "The effectiveness of energy feedback for conservation and peak demand : a literature review," Open Journal of Energy Efficiency, vol. 2, no. 1, 2013, pp. 7–15. [Online]. Available: http://eprints.qut.edu.au/58017/

[25] G. Wood and M. Newborough, "Energy-use information transfer for intelligent homes: Enabling energy conservation with central and local displays," Energy and Buildings, vol. 39, no. 4, 2007, pp. 495–503. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378778806002271

[26] T. Kim, H. Hong, and B. Magerko, "Design requirements for ambient display that supports sustainable lifestyle," in Proceedings of the 8th ACM Conference on Designing Interactive Systems, ser. DIS '10, December 2010, pp. 103–112.

[27] L. Broms et al., "Coffee maker patterns and the design of energy feedback artefacts," in Proceedings of the 8th ACM Conference on Designing Interactive Systems, ser. DIS '10, December 2010, pp. 93–102.

[28] S. Kuznetsov and E. Paulos, "UpStream: Motivating water conservation with low-cost water flow sensing and persuasive displays," in Proceedings of the 28th international conference on Human factors in computing systems, ser. CHI '10, April 2010, pp. 1851–1860.

[29] M. Laschke, M. Hassenzahl, S. Diefenbach, and M. Tippkämper, "With a little help from a friend: A shower calendar to save water," in CHI '11 Extended Abstracts on Human Factors in Computing Systems, ser. CHI EA '11, May 2011, pp. 633–646.

[30] J. Ham, C. Midden, and F. Beute, "Can ambient persuasive technology persuade unconsciously? Using subliminal feedback to influence energy consumption ratings of household appliances," in Proceedings of the 4th International Conference on Persuasive Technology, ser. Persuasive '09, April 2009, pp. 29:1–29:6.

[31] A. Gustafsson, M. Bång, and M. Svahn, "Power explorer: A casual game style for encouraging long term behavior change among teenagers," in Proceedings of the International Conference on Advances in Computer Enterntainment Technology, ser. ACE '09, October 2009, pp. 182–189.

[32] J. Mankoff, D. Matthews, S. R. Fussell, and M. Johnson, "Leveraging social networks to motivate individuals to reduce their ecological footprints," in Proceedings of the 40th Hawaii International Conference on System Sciences, ser. HICSS 2007, January 2007, pp. 87–96.

[33] D. Foster, S. Lawson, M. Blythe, and P. Cairns, "Wattsup?: Motivating reductions in domestic energy consumption using social networks," in Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries, ser. NordiCHI '10, October 2010, pp. 178–187.

[34] P. Petkov, F. Köbler, M. Foth, and H. Krcmar, "Motivating domestic energy conservation through comparative, community-based feedback in mobile and social media," in Proceedings of the 5th International Conference on Communities and Technologies, ser. C&T '11, June 2011, pp. 21–30.

[35] C. Midden and J. Ham, "The persuasive effects of positive and negative social feedback from an embodied agent on energy conservation behavior," in Proceedings of the AISB 2008 Symposium on Persuasive Technology, ser. AISB 2008, vol. 3, April 2008, pp. 9–13.

[36] M. Kugler, F. Reinhart, K. Schlieper, M. Masoodian, B. Rogers, E. André, and T. Rist, "Architecture of a ubiquitous smart energy management system for residential homes," in Proceedings of the 12th Annual Conference of the New Zealand Chapter of the ACM Special Interest Group on Computer-Human Interaction. New York, NY, USA: ACM, 2011, pp. 101–104.

[37] JSON-RPC Working Group, "JSON-RPC 2.0 Specification," http://www.jsonrpc.org/spec.html, [retrieved: December, 2014].

[38] R. T. Fieldings, "Representational State Transfer (REST). architectural styles and the design of network-based software architectures," Ph.D. dissertation, University of California, Irvine, 2000.

[39] Oracle Corporation, "Java Persistence API 2.0 Specification," http://jcp.org/aboutJava/communityprocess/final/jsr317/index.html, [retrieved: December, 2014].

[40] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley Professional, 1994.

[41] JBoss Drools, "Drools – the business logic integration platform," http://drools.jboss.org/, [retrieved: December, 2014].

[42] UserLand Software, Inc., "XML-RPC specification," http://xmlrpc.scripting.com/spec.html, [retrieved: December, 2014].

[43] M. Kugler, E. André, M. Masoodian, F. Reinhart, B. Rogers, and K. Schlieper, "Assisting inhabitants of residential homes with management of their energy consumption," in Proceedings of the The 4th International Conference on Sustainability in Energy and Buildings, ser. KES series in Smart Innovation, Systems and Technologies, vol. 22. Springer Verlag, 2013, pp. 147–156.

[44] E. Tufte, "Slopegraphs for comparing gradients: Slopegraph theory and practice," http://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0003nk&topic_id=1, [retrieved: December, 2014].

[45] C. Park, "Edward Tufte's "Slopegraphs"," http://charliepark.org/slopegraphs, [retrieved: December, 2014].

[46] S. Wendzel, T. Rist, E. André, and M. Masoodian, "A secure interoperable architecture for building-automation applications," in Proceedings of the 4th International Symposium on Applied Sciences in Biomedical and Communication Technologies, ser. ISABEL '11. New York, NY, USA: ACM, 2011, pp. 8:1–8:5. [Online]. Available: http://doi.acm.org/10.1145/2093698.2093706

[47] B. J. Fogg, "Captology: The study of computers as persuasive technologies," in CHI '97 Extended Abstracts on Human Factors in Computing Systems: Looking to the Future, ser. CHI EA '97, March 1997, p. 129.

[48] ——, Persuasive Technology: Using Computers to Change What We Think and Do. Morgan Kaufmann, 2003.

[49] P. Hartman, J. P. Bezos, S. Kaphan, and J. Spiegel, U.S. Patent 5960411: Method and system for placing a purchase order via a communications network. Amazon.com, Inc., September 1999.

Security System for Connected End-point Devices in a Smart Grid with Commodity Hardware

Hiroshi Isozaki[1, 2], Jun Kanai[1]
[1] Corporate R&D Center, Toshiba Corporation,
Kawasaki, Kanagawa, Japan
{hiroshi.isozaki, jun.kanai}@toshiba.co.jp
[2] Graduate School of Media and Governance,
Keio University,
Fujisawa, Kanagawa, Japan

Shunsuke Sasaki, and Shintarou Sano
Center for Semiconductor Research & Development,
Toshiba Corporation,
Kawasaki, Kanagawa, Japan
{shunsuke.sasaki, shintarou.sano}@toshiba.co.jp

Abstract— **Security has an important bearing on achieving successful commercial deployment of smart grids. In particular, availability is accorded the highest priority in smart grids. For end-point devices, such as smart meters or concentrators, this must be true since they must always be working. We present LiSTEE^{TM} Recovery, an architecture for a fault-tolerant system enabling end-point devices to monitor the status of the operating system and to recover even if they stop working owing to unexpected behavior or cyber-attacks, including zero-day attacks. LiSTEE^{TM} Recovery provides further functions to prevent illegitimate memory modification and to notify a head-end system once a security incident occurs. We demonstrate a full implementation of LiSTEE^{TM} Recovery on a TrustZone-capable ARM-based processor. Our experiment shows that the performance degradation is sufficiently small to be ignored. Furthermore, we observed that the cost of production and maintenance can be minimized.**

Keywords-Smart Grid, Smart Meter, Concentrator, Security, High Availability, TrustZone

I. INTRODUCTION

This paper presents a security system for connected end-point devices in smart grids. It proposes an architecture for a secure fault-tolerant system with commodity hardware and presents a detailed perspective on earlier work by the same authors [1]. In smart grids, requirements for the support of various protocols and functions to network connected end-point devices, such as smart meters or concentrators, make their systems more complicated. Because a large quantity of source code is generally necessary to implement a complicated system, the risk of including vulnerability in the system increases. Moreover, since the devices are connected to home networks, the risk of devices being attacked is high compared with legacy devices connected only to a managed network. In fact, it is reported that smart meters from various vendors were found to improperly handle malformed requests that could be exploited to cause buffer overflow vulnerability; allowing an attacker to cause a system to become unstable or freeze [2]. To keep devices secure in this situation, many security protocols and algorithms have been proposed to securely distribute a shared key between devices and head-end systems or to store privacy data in devices in a secure manner [3][4]. However, confidentiality and integrity are insufficient to solve the security problem in smart grids. Keeping high availability of the devices is strongly desired since they must always be working to provide demand-response services or to use consumption data for payment [5][6]. As a single vulnerability may cause the system to go down, it is very difficult to keep high availability in a complicated system. Furthermore, unlike in the case of interactive devices, such as PCs or smartphones, it is unreasonable to expect end users to reset and restart devices once they freeze or hang since end users cannot recognize the status of the devices and cannot determine whether the device should be rebooted or not. Thus, how to keep the availability of the devices in smart grids is a significant challenge.

To address these problems, we propose LiSTEE^{TM} Recovery, an architecture for fault-tolerant systems that automatically recovers from error status. To achieve this goal, LiSTEE^{TM} Recovery isolates a surveillance process observing the state of the system and a recovery process that reboots the system when it detects the system freezes. In LiSTEE^{TM} Recovery, surveillance and recovery processes run in an isolated secure environment whereas general-purpose processes, including the operating system, such as network or storage access, run in a non-secure environment with hardware access control performed with respect to memory. Hence, a memory area where surveillance and recovery processes are arranged cannot be accessed by general-purpose processes. As a result, even if the operating system is attacked and crashes, it becomes possible to prevent interference in the surveillance and recovery processes.

The remainder of this paper is organized as follows. In Section II, problems are defined. Section III presents background information. Sections IV and V propose a framework and implementation of LiSTEE^{TM} Recovery. The evaluation of LiSTEE^{TM} Recovery is shown in Section VI, related work is referred to in Section VII, and the paper concludes with Section VIII, which is devoted to the conclusion and future work.

II. PROBLEM DEFINITION

In a legacy system, surveillance and recovery processes and their execution environment are monolithically configured. In other words, the reliability of surveillance and recovery processes depends on the reliability of their execution environment. In order to keep reliability high, a

system needs to be implemented without vulnerability. In order to detect and eliminate vulnerability in source code, various testing methods have been proposed [7][8]. However, since end-point devices will be deployed without maintenance over a long period of time within smart grids and new vulnerabilities are found day after day, there is a large risk that such devices will continue operating without vulnerabilities being fixed even if those devices had no vulnerabilities at the time of shipping. For example, there is a well-known attack against x86 processors called "Ret2Libc" which enables an attacker to inject and execute code, and it had been regarded as invalid against ARM processors [9]. However, once new attack which is similar with Ret2Libc against ARM processors has been proposed, buffer over flow on ARM processor has been regarded as real theat. Therefore, attackers may exploit a vulnerability, such as buffer overflow or malformed network input, in order to cause the device to crash. To make matters worse, attackers are in a somewhat advantageous position in launching a large attack since the number of device vendors is limited and the software installed in the devices is uniform. Furthermore, attackers can reverse-engineer code without administrators noticing in order to find a vulnerability since, unlike a server application, devices are located at the user side. Therefore, when attackers find one vulnerability in a single device, they can exploit it on many devices. Considering the above situation, the following problems are to be solved in order to keep high availability under a legacy system.

A. *Difficult to Keep a High Level of Surveillance Continuity*

End-point devices need to support various network protocols and data formats depending on countries or use cases in smart grids [10][11][12]. In order to minimize the implementation cost of a complicated application program or a minor network protocol on end-point devices, Linux will be used as a software execution environment. In Linux, the surveillance and recovery processes can be implemented as a user task executed on the operating system or as an interrupt handler in the operating system. When a surveillance target process is implemented as a user task running on the operating system then support functions in the operating system, such as the "cron" service in Linux, can be used to detect a failure of the user task and to automatically restart the target process. When the surveillance process is implemented as an interrupt handler in the operating system, then more sophisticated implementation is necessary than for an application program; it is automatically and periodically called by a timer interrupt as long as the operating system works. Another legacy approach is implementation of a monitoring and detecting mechanism in the operating system. For example, in order to find buffer overflow attacks, an anomaly detection method is proposed where a protection element monitors system call frequencies, and if the frequencies are different from normal behavior, it determines that an attack occurs [13]. However, the fundamental problem of a legacy approach is that there is no way to restart the process if the operating system itself crashes for any reason. Furthermore, the protection mechanism itself could be a target of the attack, and as a result the protection mechanism could be invalidated. Thus, there is a large risk of devices in a smart grid breaking down and the attack may be able to cause an extensive blackout in the worst case. In order to prevent devices breaking down, a robust method of recovering the system from failure is required in order to keep a high level of availability. Still, some existing hardware devices support a watchdog timer function that detects the status of the operating system and automatically reboots the system [14]. Since not all devices support the function and it is difficult to implement complicated functions in the system as discussed below, a new approach is desired. To clarify the conditions, only a software failure including an attack is assumed in this paper. A physical fault, such as a hardware failure or loss of power, or a hardware attack, such as physically destroying devices or cutting cables, are beyond the scope of this paper.

B. *Difficult for an Administrator to Detect when an Incident Occurs*

End-point devices are connected with a head-end system through the network to provide demand-response services. When the devices detect an error status, such as a surveillance target process being stopped for an unknown reason, it is desirable for these devices to send a report to the head-end system so that an administrator can realize the situation and use the report to investigate the reason for the failure. However, for the reason described above, there is no way for devices to send a message to the head-end system if the operating system crashes in a system where the network connectivity function is implemented as a user task or it is implemented within the operating system. Even in such a case, it is desirable to provide a method enabling devices to send a message to acknowledge the error situation to the system administrator. In addition to the unexpected failure, attacks on the network connectivity function need to be considered. When an attacker gains full access to the system under control, the attacker may try to disable the network connectivity function in the operating system. Therefore, it is desired not simply to provide a method of sending a message but to keep the network connectivity function secure to protect it against the attack even if the operating system is modified or the control of the operating system is taken over.

Besides notification of the error situation to the system administrator, a software update function is also desirable. However, since many existing hardware devices already support a secure firmware update function and its method is highly dependent on each device, it is beyond the scope of this paper.

In addition to the problem described above, the following business problem needs to be considered when introducing a new architecture to the market.

C. *Development and Production Cost*

Cost is an important aspect in evaluating the proposed security architecture. Generally, there are two types of cost: development cost, consisting primarily of personnel expenses, and production cost, which is charged per device.

When implementing an end-point device, if the new security architecture requires a complete software rebuild, the architecture will never be commercialized. Thus, it is desirable to reuse existing software assets, such as libraries, middleware and applications, as much as possible in order to minimize the development cost, including the verification cost. In the case of smart grids, the verification cost is large since reliability is strongly required. Besides the development cost, we need to consider the cost per device. One approach to solve the problems described above is to utilize a dedicated hardware security chip. However, since such chips tend to be very expensive, their use may raise production cost per device. Therefore, the use of widely available existing commodity hardware is desirable in order to minimize production cost.

III. BACKGROUND (TRUSTZONE)

In this section, we provide background information on the hardware technologies leveraged by LiSTEE^{TM} Recovery.

A. ARMv7 Architecture

ARM processors support different processor modes depending on the architecture version. The ARMv7 architecture on which LiSTEE^{TM} Recovery is implemented supports the seven processor modes shown in Table I.

TABLE I. ARM PROCESSOR MODE AND BANK REGISTER

Mode	*level*	*description*	*Bank register*	*# of bank registers*
USR	unprivileged	User mode	r8-r14	7
SVC	privileged	Supervisor mode	r13-r14, spsr	3
IRQ	privileged	IRQ mode	r13-r14, spsr	3
FIQ	privileged	FIQ mode	r8-r14, spsr	8
ABT	privileged	Abort mode	r13-r14, spsr	3
UND	privileged	Undefined mode	r13-r14, spsr	3
MON	privileged	Monitor mode	r13-r14, spsr	3

The processor is executed by selectively switching the modes depending on the process. The processor mode is changed either when a program, such as an operating system, calls a dedicated instruction or when software or hardware exception occurs. The seven modes are categorized as either non-privileged mode or privileged mode by privilege level. In a general system, an operating system is executed in privileged mode and application programs are executed in unprivileged mode. In privileged mode, execution of all instructions and access to all memory regions are allowed, whereas in unprivileged mode availability of instructions and accessibility of memory regions are restricted.

The ARMv7 processor has 40 registers, consisting of 33 general registers and 7 status registers. These registers are arranged in partially overlapping banks. For example, r13, which is a bank register and usually used for stack pointer, refers to different physical registers in User mode and Supervisor mode. For non-banked registers, which refer to the same physical register in different modes, an operating system needs to save and restore in working memory when switching from one mode to another mode so that execution can be subsequently resumed from the same point. On the contrary, the operating system does not need to save the context of banked registers. For example, the operating system does not need to save the context of r13 when switching from User mode to Supervisor mode. Therefore, rapid context switching is enabled.

B. TrustZone

TrustZone is a hardware security function supported by a part of the ARM processor [15][16]. In addition to unprivileged mode and privileged mode, a TrustZone-enabled ARM processor supports two worlds that are independent of the modes. One is the secure world for the security process and the other is the non-secure world for everything else. Each processor mode shown in Table I is available in both the secure world and the non-secure world. Fig. 1 shows the relationship between worlds and modes conceptually. The world in which the processor is executing is indicated by the NS-bit in the Secure Configuration Register (SCR) except when the processor is in monitor mode. When the processor is in monitor mode, it is in the secure world regardless of the value of the NS-bit of SCR. The processor is executed by selectively switching the worlds if necessary. For example, it is assumed that the key calculation process is executed in the secure world and all other general processes, such as storage access or network accesses are executed in the non-secure world.

The software that manages switching between the secure world and the non-secure world is called the monitor. The monitor is executed in monitor mode. TrustZone provides a dedicated instruction, the Secure Monitor Call (SMC) instruction, to transit between the worlds. As soon as the SMC instruction is called, the processor switches to monitor mode. Monitor saves a context of the program running in the current world on the memory and restores a context of the program running in the previous world, then changes the world to set the NS-bit of SCR, and finally executes the program running in the previous world. Besides the SMC

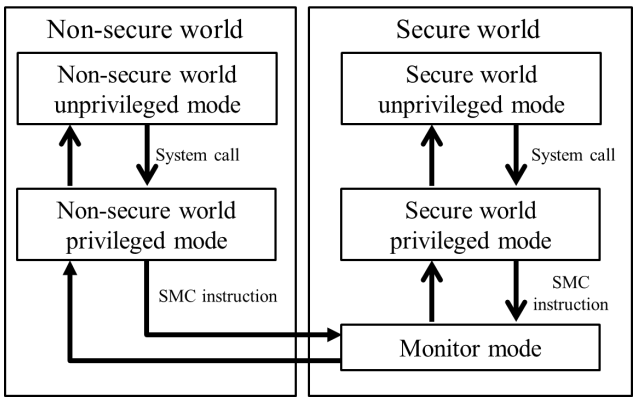


Figure 1. Mode and world in ARM.

instruction, hardware exceptions can be configured to cause the processor to switch to monitor mode.

Note that general registers and Saved Program Status Register (spsr) are not banked between worlds. For example, when r13 in User mode of the secure world is referred and the monitor switches from the secure world to the non-secure world, and then r13 in User mode of the non-secure world is referred, the same physical register is referred. Therefore, the monitor needs to save and restore both bank registers and non-bank registers when it switches worlds.

By using TrustZone-capable hardware, it is possible to make a system where a process running in the secure world can access all system resources, such as memory or peripherals, whereas a process running in the non-secure world can access only a part of system resources. For example, when used in combination with the TrustZone Address Space Controller (TZASC), access to a particular region of working memory can be restricted for a process running in the non-secure world even if the process runs in privileged mode. When a process running in the non-secure world accesses a memory region that it is configured to be prohibited from accessing from a process running in the non-secure world, TZASC generates an interrupt signal and it is sent to the processor. As a result, the violation causes an external asynchronous abort. Similar to TZASC, when used in combination with the TrustZone Protection Controller (TZPC), access to a peripheral can be restricted for a process running in the non-secure world. In contrast to TZASC, the access control policy of TZPC can be configured per peripheral, such as DRAM, Timer, or Real-Time Clock (RTC). That is, the configuration of TZPC is performed peripheral by peripheral. There is a correlation between TZASC and TZPC. For example, when configuring a policy such that access to a particular region of DRAM is restricted, the access control of TZPC corresponding to DRAM is set to off and the proper access control policy with the corresponding region is installed on TZASC. TZPC is configured as secure when booting the system. Therefore, for all peripherals whose access controls are valid by TZPC, access by a process running in the non-secure world is prohibited by default. TZASC and TZPC can only be configured by a process running in the secure world, in order to protect those configurations from illegitimate modification.

IV. Framework of LiSTEE^{TM} Recovery

LiSTEE^{TM} Recovery provides a method for an end-point device to automatically recover from an error status. It also provides a high-level memory protection mechanism. Hence, the recovery process is securely executed without interference. Fig. 2 shows the entire architecture of LiSTEE^{TM} Recovery. LiSTEE^{TM} Recovery consists of three components: Normal OS, LiSTEE^{TM} Tracker Application (LiSTEE^{TM} TA), and LiSTEE^{TM} Monitor.

- Normal OS: An operating system that executes general-purpose processes, such as storage access or network communication. It is executed in the non-secure world. All applications implementing smart meter functions or concentrator functions run on this operating system.
- LiSTEE^{TM} Tracker Application (LiSTEE^{TM} TA): Surveillance and recovery processes executed in privileged mode in the secure world. LiSTEE^{TM} TA includes three modules: Watcher module, Recovery module, and Notification module. The Watcher module is an entry point of LiSTEE^{TM} TA. It is executed periodically by a timer interrupt through LiSTEE^{TM} Monitor. Whenever it is called, it investigates the status of Normal OS. If it detects Normal OS is not working, it calls the Recovery module to reboot the system. Otherwise, it calls the SMC instruction to switch to Normal OS. Moreover, the Notification module is called before the Recovery module reboots the system. It sends a message to notify that the system is about to reboot to the head-end system through network.
- LiSTEE^{TM} Monitor: A program running in the monitor mode. It initializes configurations of TrustZone-related hardware when booting the system. It also provides a context switching function between worlds in the hardware interrupt handler and the SMC handler. Moreover, LiSTEE^{TM} Monitor manages the access control policy and installs the policy on TZASC when booting. Policy Manager takes on their roles.

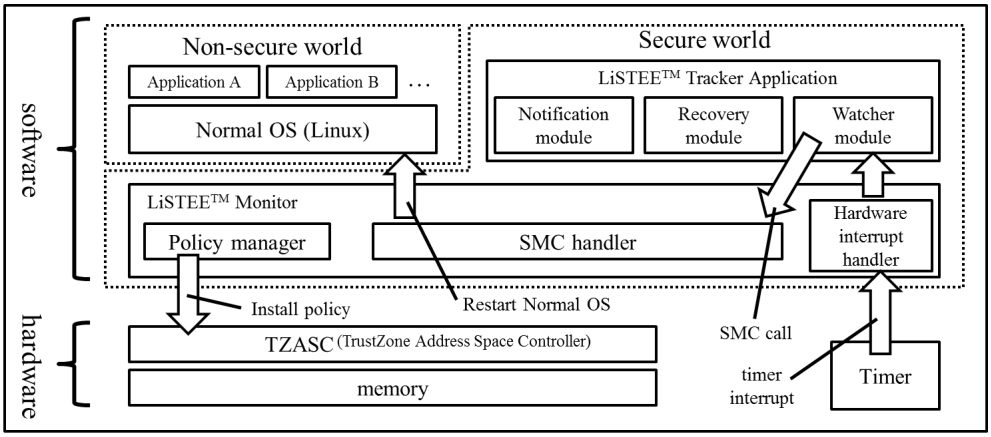


Figure 2. System Architecture of LiSTEE^{TM} Recovery.

The primary feature of LiSTEE^{TM} Recovery is to provide a method for the end-point device to detect the status of Normal OS and to recover it even if Normal OS crashes or stops working. Furthermore, it provides two additional functions. One is to enhance the security protection for LiSTEE^{TM} Monitor, LiSTEE^{TM} TA and Normal OS against attacks. The other is to send a message to the head-end system when an incident occurs. The details of these functions are described below.

A. Baseline Common Functions

LiSTEE^{TM} Monitor has the role of providing baseline common functions to operate Normal OS and LiSTEE^{TM} TA concurrently. LiSTEE^{TM} Monitor has two functions; system initialization and context switching between worlds.

1) System initialization

When booting the system, the processor is in the secure world and LiSTEE^{TM} Monitor is firstly executed. To run Normal OS and LiSTEE^{TM} TA concurrently, it needs to load and execute both of them. It first initializes the status of the processor in both worlds, and loads LiSTEE^{TM} TA in the secure world. Then, it invokes context switching to transit from the secure world to the non-secure world, loads the boot loader program of Normal OS, and executes it in the non-secure world. Finally, the boot loader program loads Normal OS and executes it.

The TrustZone-enabled processor supports the function that is either monitor or Normal OS traps each exception (IRQ, FIQ, and external abort). When booting the system, LiSTEE^{TM} Monitor configures that hardware interrupt handler in LiSTEE^{TM} Monitor traps timer interrupt so that Normal OS cannot interfere with the execution of LiSTEE^{TM} TA when timer interrupt occurs. As well as timer interrupt, LiSTEE^{TM} Monitor configures that hardware interrupt handler in LiSTEE^{TM} Monitor traps external abort. Since the access violation causes external abort as described above, this configuration enables LiSTEE^{TM} TA to detect the occurrence of a memory access violation.

TZPC is configured to be accessed from the secure world only when booting the system. Since Normal OS needs to use peripherals, LiSTEE^{TM} Monitor needs to change the configuration of TZPC to non-secure. The only exception is Timer, which triggers periodical execution of LiSTEE^{TM} TA. Since it is necessary to prevent the configuration of Timer from changing by a process running in the non-secure world, LiSTEE^{TM} Monitor remains the configuration of TZPC corresponding to Timer as secure.

2) Context Switching between Worlds

In LiSTEE^{TM} Recovery, the trigger of context switching between worlds is either the SMC instruction or the Timer interrupt caused by the hardware timer. The SMC handler in LiSTEE^{TM} Monitor is executed when the SMC instruction is called and it transits from the secure world to the non-secure world. In contrast to the SMC handler, the timer interrupt triggers transit from the non-secure world to the secure world. In both cases, LiSTEE^{TM} Monitor invokes context switching between worlds. It first determines the current world. As described in section III-B, general registers and Saved Program Status Register are not banked between worlds. Therefore, LiSTEE^{TM} Monitor needs to save the contents of the registers belonging to the current world on working memory to prevent loss of the previous context, and then change the world. Finally, it restores the contents of the registers belonging to the transition destination world and resumes the execution.

B. Periodical Surveillance and Recovery

While executing Normal OS, whenever the timer interrupt occurs, the processor jumps to the hardware interrupt handler in LiSTEE^{TM} Monitor. The hardware interrupt handler context switches from the non-secure world to the secure world and calls LiSTEE^{TM} TA. Specifically LiSTEE^{TM} Monitor saves a context of Normal OS to memory and restores a context of LiSTEE^{TM} TA, then changes the world and finally calls the Watcher module of LiSTEE^{TM} TA. The Watcher module checks the status of Normal OS. If it judges that Normal OS is not working, the Watcher module calls the Recovery module that reboots the system. Otherwise, it calls the SMC instruction. Then, the SMC handler in the LiSTEE^{TM} Monitor is executed. It context switches from LiSTEE^{TM} TA to Normal OS, and restarts Normal OS at the point just before the timer interrupt occurred. While executing LiSTEE^{TM} Monitor and LiSTEE^{TM} TA, the execution of Normal OS is suspended. That is, Normal OS continues to be processed as if nothing were executed during the execution of LiSTEE^{TM} TA. Fig. 3 shows the flowchart of the periodic surveillance and recovery process.

There are many ways for the Watcher module to determine whether Normal OS is working or not. One of the methods is to check the data area of Normal OS. In general, when an operating system is working, there must be a certain data area that is updated regularly. By checking this data area, it is possible for the Watcher module to judge whether Normal OS is working or not.

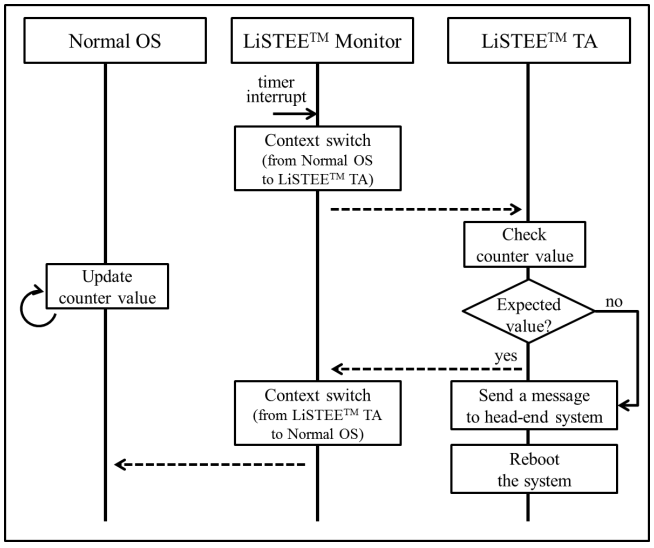


Figure 3. Flowchart of periodical surveillance and recovery.

C. Memory Protection

By utilizing TZASC, LiSTEE^{TM} Monitor provides an access control function such that access of Normal OS running in non-secure mode to the working memory area, which LiSTEE^{TM} Tracker Application running in the secure world uses, is subject to restrictions. Policy Manager in LiSTEE^{TM} Monitor manages three kinds of access control policies: full access, access denied, and read-only. When booting the system, Policy Manager divides working memory into several regions and it installs one of the three access control policies for each working memory region on TZASC before loading Normal OS.

Table II shows how each policy works. Full access indicates no restriction. A process running in both non-secure world and secure world can freely access the region configured according to this policy. This policy is primarily used to share data between Normal OS and LiSTEE^{TM} TA. Access denied indicates full restriction. A process running in the non-secure world can neither read nor write to a region configured according to this policy, whereas a process running in secure world can read and write to the region. Read-only indicates a process running in the non-secure world cannot overwrite the content on the memory, whereas a process running in the secure world can freely access the region using ordinary random access memory, such as DRAM or SRAM, as the working memory which is, of course, physically writable memory.

TABLE II. ACCESS CONTROL POLICY

Policy	*From secure world process*	*From non-secure world process*	
		Read	Write
Full access	OK	OK	OK
Access denied	OK	NG	NG
Read-only	OK	OK	NG

Using these policies, LiSTEE^{TM} Recovery provides two memory protection mechanisms. Fig. 4 shows how these memory protection mechanisms work. One is protection for the kernel area of Normal OS. The other mechanism is protection for LiSTEE^{TM} Monitor and LiSTEE^{TM} TA.

To realize protection for the kernel area of Normal OS, LiSTEE^{TM} Monitor provides read-only memory. In general, when a program is loaded into memory, a data region (data segment) and a code region (code segment) are assigned. In the initial state before booting the system, all regions are allowed to be accessed from the non-secure world by default. In order to allow the boot loader to write the code segment into the memory, LiSTEE^{TM} Monitor leaves the memory region as is until the code segment is loaded. Just after executing the kernel of Normal OS, LiSTEE^{TM} Monitor sets the memory region as read-only for kernel code segment of Normal OS. As a result, even Normal OS is prohibited from overwriting its own code segment.

To protect LiSTEE^{TM} Monitor and LiSTEE^{TM} TA, Policy Manager in LiSTEE^{TM} Monitor installs an access control policy such that Normal OS cannot access the memory area allocated to LiSTEE^{TM} Monitor and LiSTEE^{TM} TA, whereas LiSTEE^{TM} TA and LiSTEE^{TM} Monitor can access all areas when booting the system. This policy protects LiSTEE^{TM} Monitor and LiSTEE^{TM} TA from illegitimate falsification by Normal OS, even if Normal OS is attacked and under the control of an attacker.

Besides the protection for LiSTEE^{TM} Monitor and LiSTEE^{TM} TA, memory protection provides a hardware access control mechanism. One of the possible attacks to disable end-point devices is that of shutting down the system. To prevent such an attack, Policy Manager in LiSTEE^{TM} Monitor installs an access control policy so that Normal OS cannot access the registers corresponding to power management. Thus, it is possible to protect the system against the shutdown attack even if Normal OS is under the control of an attacker.

In the case of policy configured to access denied or read-only, TZASC generates an interrupt signal when the access violation caused by a process running in the non-secure world occurs. LiSTEE^{TM} Monitor configures the hardware interrupt handler in LiSTEE^{TM} Monitor to trap the interrupt so that the system will continues to work without crashing even if access violation occurs, and LiSTEE^{TM} Monitor can detect the access violation.

D. Message Notification

LiSTEE^{TM} Recovery provides a function to notify the head-end system that Normal OS has stopped working and is rebooting the system by sending a message through the

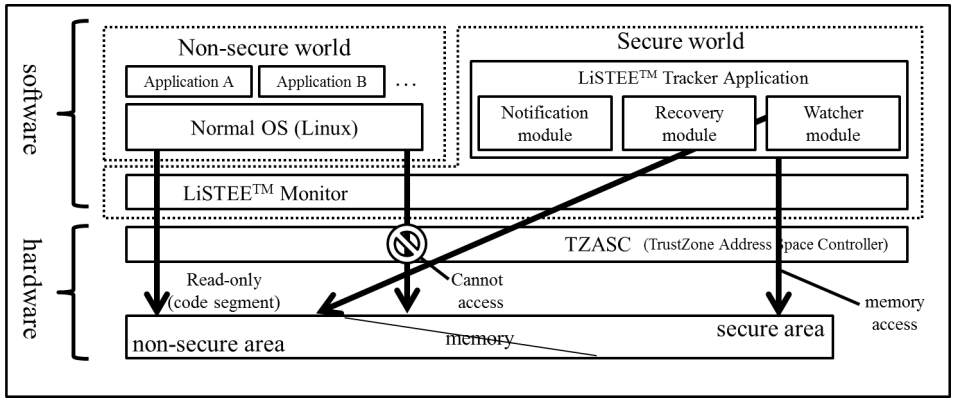


Figure 4. Memory protection mechanism.

network even if the operating system is modified or the control of the operating system is taken over; the resulting network function is disabled by an attacker. The Notification module has the role of sending a message. Although Normal OS has a network connectivity function, such as TCP/IP stack, LiSTEE^{TM} TA cannot use the function since there is a case where it is not working when sending a message. Thus, LiSTEE^{TM} TA supports the network connectivity function including the network application, the network protocol stack and the network driver to notify the error situation to the system administrator through the network. Obviously, it is possible to send a head-end system a message whenever LiSTEE^{TM} TA is executed to notify that the system works correctly.

V. PROTOTYPE IMPLEMENTATION

We used ARM C/C++ Compiler 5.01 to build LiSTEE^{TM} Monitor and LiSTEE^{TM} TA. We used gcc 4.4.1 to build Linux 3.6.1 as Normal OS. We chose Motherboard Express uATX with the CoreTile Express A9x4 processor that supports TrustZone as an execution environment.

Regarding a memory map, from 0x48000000 through 0x4A000000 is assigned for SRAM, and from 0x60000000 through 0xE0000000 is assigned for DRAM. Table III shows the memory map with the access control policy of the memory. In Table III, Normal OS (code) indicates the Linux kernel code. Normal OS (data) includes the Linux data, the application code and the application data. For clarification, full access is applied from the non-secure world for an area not described in Table III.

TABLE III. MEMORY MAP

Data	Start Address	Size	Security Permission (From non-secure world)
Vector tables + Initialization code + LiSTEE^{TM} Monitor + LiSTEE^{TM} TA	0x48000000	0x01B00000	Access denied
Normal OS (code)	0x60000000	0x002FE000	Read-Only
Normal OS (data)	0x602FE000	0x3EF02000	Full access
Shared memory	0x9F200000	0x00C00000	Full access

For the Policy Manager in LiSTEE^{TM} Monitor to install an access control policy on TZASC, the start address and the size of each memory region are predefined. After the boot loader loads Linux at the predefined value, LiSTEE^{TM} Monitor installs the access control policy on TZASC. As shown in Table III, the access to the memory regions allocated to LiSTEE^{TM} Monitor, LiSTEE^{TM} TA and the code segment of Normal OS is restricted for the Normal OS running in the non-secure world, whereas the access to the region allocated to the data segment of Normal OS and shared memory is not. For clarification, LiSTEE^{TM} Monitor and LiSTEE^{TM} TA running in the secure world can access all regions. Furthermore, since LiSTEE^{TM} Monitor sets the configuration registers of TZASC to prohibit Normal OS from accessing them, Normal OS cannot change this configuration.

Table IV shows the configuration of TZASC. In Table IV, the meaning of the value of the security permissions field is as follows: 0b1111 indicates full access from both the secure world and the non-secure world, 0b1100 indicates secure read/write is permitted but non-secure read/write is restricted (access denied), and 0b1110 indicates secure read/write and non-secure read are permitted but non-secure write is restricted (read-only). An entry with larger entry number is accorded higher priority than one with smaller entry number. Therefore, we first set all regions with a policy of full access as entry number 0, and then set access control policies from entry number 1 through 7. The size of a region to which access control is applied is discrete, such as 32 KB, 64 KB, …, 1 MB, 2 MB, 4 MB, …, 2 GB, 4 GB. Therefore, to set policy for LiSTEE^{TM} Monitor and LiSTEE^{TM} TA whose size is 0x01B00000 (27 MB), we used four entries: entry number 1 (16 MB), entry number 2 (8 MB), entry number 3 (2 MB), and entry number 4 (1 MB). In contrast to the size of LiSTEE^{TM} Monitor and LiSTEE^{TM} TA, the size of Normal OS (code) is a fraction (32 MB – 8 KB), and TZASC has restrictions such that it is impossible to define an entry whose size is smaller than 32 KB. Instead, it is possible to define a subregion to equally divide a region into eight with the access control policy, and enable the policy for each subregion. For example, when the size of a region is 32 KB, it is possible to enable a policy for each 4 KB subregion. An 8 bit subregion disable field controls enabling and disabling the policy. Each bit in a subregion disable field enables the corresponding subregion to be disabled. For example, when zero is set to the value of the highest bit in a subregion disable field, the policy for subregion 0 (the subregion having the highest address) is enabled. To set the policy for a Normal OS (code) region, we first defined two regions, 2 MB (entry number 5) and 1 MB (entry number 6) and set the read-only policy. Then, we defined the region with a size of 64 KB (entry number 7) that overlaps the last portion of entry number 6, equally divides the region into eight, sets the policy of full access, and enables the policy for the last subregion only. As a result, the policy of full access is set to the subregion having the highest address only, and the policy of read-only remains for the rest of the subregions.

As shown in Table III and Table VI, the policies can be clearly defined and there is no overlapped region. Thus, no policy conflict exists in LiSTEE^{TM} Recovery.

TABLE IV. CONFIGURATION OF TZASC

Entry Number	Start Address	Size	Subregion disable	Security Permission
0	--	--	--	0b1111
1	0x48000000	0x17(16MB)	0x0	0b1100
2	0x49000000	0x16(8MB)	0x0	0b1100
3	0x49800000	0x14(2MB)	0x0	0b1100
4	0x49A00000	0x13(1MB)	0x0	0b1100
5	0x60000000	0x14(2MB)	0x0	0b1110
6	0x60200000	0x13(1MB)	0x0	0b1110
7	0x602F0000	0xF(64KB)	0x7F	0b1111

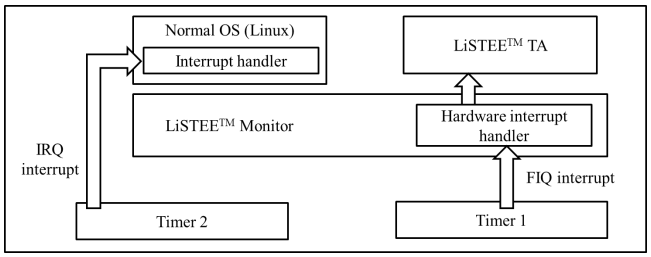


Figure 5. Assignment of timer interrupt.

Fig. 5 shows the assignment of the timer interrupt. We allocated a timer interrupt caused by a timer (timer 1) to Fast Interrupt Request (FIQ) and the timer interval was set to 1 second. The FIQ interrupt is handled by the hardware interrupt handler in LiSTEE™ Monitor, then it calls LiSTEE™ TA and, as a result, LiSTEE™ TA is periodically called. We used another timer (timer 2) and allocated it to Interrupt Request (IRQ), and the timer interval was set to 4 milliseconds. The IRQ interrupt is handled by the interrupt handler in Linux. Since Linux assumes the timer interrupt is allocated to IRQ, modification of the Linux source code to adopt LiSTEE™ Monitor is unnecessary.

Table V shows a configuration of hardware interrupt. We configured Secure Configuration Register (SCR) and Current Program Status Register (CPSR) so that the FIQ handler of LiSTEE™ Monitor is called when the FIQ interrupt occurs, whereas the IRQ handler in Linux is called when the IRQ interrupt occurs during executing Linux. Table VI shows the register setting to achieve the configuration of Table V. CPSR.I indicates the Interrupt disable bit and is used to mask the IRQ interrupt. CPSR.F indicates the Fast interrupt disable bit and is used to mask the FIQ interrupt. CPSR.A indicates the asynchronous abort disable bit and is used to mask asynchronous abort. SCR.FIQ controls which mode the processor enters when the FIQ interrupt occurs. If one is set, it enters monitor mode, otherwise it enters FIQ mode. SCR.IRQ controls which mode the processor enters when the IRQ interrupt occurs. If one is set, it enters monitor mode, otherwise it enters IRQ mode. SCR.FW controls whether the F bit in the CPSR can be modified in the non-secure world. SCR.EA controls which mode the processor enters when external abort including the one generated by TZASC. If one is set, it enters monitor mode, otherwise it enters abort mode. SCR.AW controls whether the A bit in the CPSR can be modified in the non-secure world. If zero is set, CPSR.A can be modified only in the secure world, otherwise it can be modified in both worlds.

TABLE V. RELATIONSHIP BETWEEN WORLD AND INTERRUPT

World when interrupt occurs	Interrupt	Jumps to
Non-secure world	FIQ	Hardware interrupt handler (FIQ handler) in LiSTEE™ Monitor
	IRQ	IRQ handler in Normal OS (Linux)
Secure world	FIQ	Pending FIQ
	IRQ	Pending IRQ

TABLE VI. CPSR AND SCR REGISTER SETTING

		Non-secure world	Secure world (LiSTEE™ TA)	Secure world (Monitor)
CPSR	I	0/1 (depending on the configuration of Normal OS)	1 (IRQ disabled)	1 (IRQ disabled)
	F	0 (FIQ enabled)	1 (FIQ disabled)	1 (FIQ disabled)
	A	0 (Asynchronous abort enabled)	0 (Asynchronous abort enabled)	1 (Asynchronous abort disabled)
SCR	FIQ	1 (enter monitor mode)	0 (enter FIQ mode)	0/1 (depending on which world transiting to)
	IRQ	0 (enter IRQ mode)	0 (enter IRQ mode)	0 (enter IRQ mode)
	FW	0 (can be modified CPSR.F only in secure)	0 (can be modified CPSR.F only in secure)	0 (can be modified CPRS.F only in secure)
	EA	1 (enter monitor mode)	0 (enter abort mode)	0/1 (depending on which world transiting to)
	AW	0 (can be modified CPSR.A only in secure)	0 (can be modified CPSR.A only in secure)	0 (can be modified CPSR.A only in secure)

As shown in Table V, when a processor is in the non-secure world and the FIQ interrupt assigned for timer 1 occurs, the FIQ handler in monitor mode is called since one is set to SCR.FIQ. The FIQ handler in monitor mode switches from the non-secure world to the secure world and calls the FIQ handler in LiSTEE™ TA. Finally, the FIQ handler in LiSTEE™ TA calls the Watcher module. The entry point to LiSTEE™ TA from LiSTEE™ Monitor is only the FIQ handler in LiSTEE™ TA and it never returns to LiSTEE™ TA after the Watcher module calls SMC instruction under the current implementation. When considering returning to the original location in LiSTEE™ TA when entering the secure world next time as future extension, the FIQ handler in monitor mode sets the instruction located in the address next to the address of the instruction just after calling the SMC instruction in the previous time to r14 before calling the FIQ handler of LiSTEE™ TA. On the other hand, when the IRQ interrupt occurs, the IRQ handler in Normal OS is called. Furthermore, Normal OS cannot change the configuration of CPSR.F since zero is set to SCR.FW. Therefore, the FIQ interrupt is always enabled and the timer interrupt is input to the monitor.

When a processor is in the secure world and FIQ or IRQ interrupt occurs, the interrupt is pending since zero is set to CPSR.F and CPSR.I. For future extension, LiSTEE™ Monitor changes SCR.FIQ setting during context switching so that LiSTEE™ TA handles the FIQ interrupt directly without LiSTEE™ Monitor when the FIQ interrupt occurs in the secure world. That is, zero is set to SCR.FIQ when it transits from the non-secure world to the secure world to jump to the FIQ handler in LiSTEE™ TA when the FIQ interrupt occurs in the secure world. On the other hand, one is set when it transits from the secure world to the non-secure world to enter monitor mode when the FIQ interrupt occurs in the non-secure world.

When a processor is in monitor mode, FIQ and IRQ interrupt are disabled to avoid occurrence of multiple interrupt.

In order to determine whether Linux is working or not, we made a small application program, which runs on Linux and communicates with LiSTEE^{TM} TA. Shared memory is used to exchange data between LiSTEE^{TM} TA and Normal OS. The application program writes a counter value into the shared memory periodically. Then LiSTEE^{TM} TA reads the counter value from the shared memory. When Normal OS is crashed, the application program cannot update the counter value. If the counter value is not updated in a certain amount of time or the counter value is not an expected value, LiSTEE^{TM} TA determines that Normal OS is not working. Another method of checking the status of Normal OS is to monitor the status of a specific field, such as a task structure or page tables in Normal OS, but we have not implemented it. Thanks to the memory protection function, it is impossible for Normal OS to check the checking process running in LiSTEE^{TM} TA. Since it is possible to maintain secrecy of Normal OS as to which memory area of Normal OS LiSTEE^{TM} TA monitors or how often LiSTEE^{TM} TA checks it, it is difficult for an attacker to plan a countermeasure to circumvent the checking.

LiSTEE^{TM} Recovery provides a method to continue working even if a memory access violation caused by TZASC occurs. Fig. 6 shows the flowchart of how LiSTEE^{TM} TA and LiSTEE^{TM} Monitor recover from the error status to the normal status when an access violation caused by TZASC occurs. When booting the system, LiSTEE^{TM} Monitor configures SCR.EA so that external aborts including the ones TZASC generates are handled in Monitor mode, instead of by the abort handler in Normal OS. Furthermore, it is prohibited to mask external abort from the non-secure world to configure SCR.AW. Therefore, when an access violation occurs in user mode in the non-secure world, for example, a processor jumps to the abort handler in LiSTEE^{TM} Monitor. At this time, the values of r14 (lr) and spsr are the values of PC (Program Counter) and spsr of the mode just before the access violation occurs, respectively. The abort handler in LiSTEE^{TM} Monitor saves registers including r14 and spsr of original mode in the non-secure world on working memory, context switches from the non-secure world to secure world, and calls the abort handler in LiSTEE^{TM} TA. The abort handler in LiSTEE^{TM} TA checks the status of Normal OS. For example, LiSTEE^{TM} TA checks which process running in Normal OS triggers access violation or checks memory address where an access violation is triggered to investigate the reason for the access violation later. After LiSTEE^{TM} TA checks the status, it calls the SMC instruction and jumps to LiSTEE^{TM} Monitor. While LiSTEE^{TM} TA works in the background when an access violation occurs, LiSTEE^{TM} Recovery behaves as if data abort occurs from the viewpoint of Normal OS. When data abort occurs, a processor automatically stores PC and cpsr of the mode just before data abort occurs to r14 and spsr respectively. LiSTEE^{TM} Monitor carries out a similar operation with the processor when an access violation occurs. LiSTEE^{TM} Monitor switches from the secure world to the non-secure world, restores the saved values including setting the saved value of r14 and spsr just before the access violation occurs to banked registers for abort mode in order to be able to return to the original location after exiting abort mode, and calls the abort handler of Normal OS. Therefore, when Normal OS restarts a process, the data abort handler is executed.

When LiSTEE^{TM} TA determines that Linux is not working, it sends the head-end system a message. In order to send a message to the head-end system when LiSTEE^{TM} TA detects that Linux is not working, we ported a network driver and UDP/IP stack to LiSTEE^{TM} TA. We defined a proprietary protocol and data format over UDP to notify the head-end system that LiSTEE^{TM} TA starts reboot of the system. An application data size of UDP packet is 32 bytes, and it consists of 4 bytes of device ID, 1 byte of flag indicating the status of the device, and 27 bytes of reserved area.

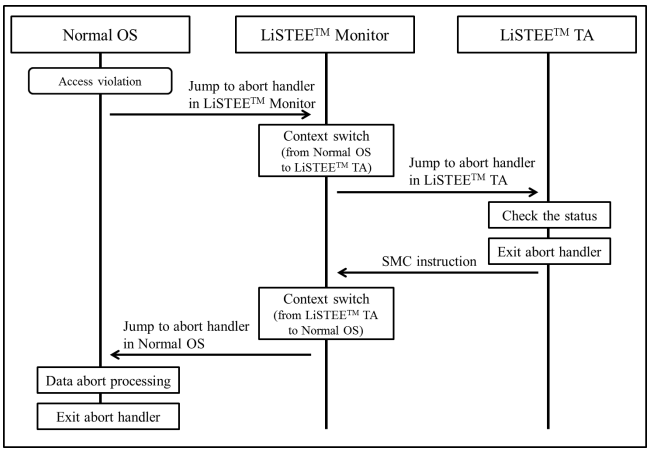


Figure 6. Flowchart of access violation handling.

VI. EVALUATION

In this section, we describe the result of the evaluation in terms of security to verify the problems of the legacy system defined in Section II can be solved. Performance and cost analysis of LiSTEE^{TM} Recovery is also described below.

A. Security Analysys

1) Surveillance and Recovery: LiSTEE^{TM} Recovery can recover from a failure to reboot the system even if Normal OS crashes. The reason for the crash could be a software bug or a cyber-attack, including a zero-day attack prompted by unknown vulnerabilities. In either case, since the hardware timer interrupt continues working regardless of the state of Normal OS, LiSTEE^{TM} TA is always periodically called and can detect a failure of Normal OS. At the next level, it is desirable to detect the failure as soon as possible. Detection time depends on how frequently LiSTEE^{TM} TA checks the status of Normal OS. Since the execution time of LiSTEE^{TM} TA and context switching by LiSTEE^{TM} Monitor is very short, LiSTEE^{TM} Recovery can detect the crash of Normal OS very quickly. Some attackers may continue to

attack just after rebooting the system. One possible approach to a countermeasure for the attack is to let LiSTEE^{TM} TA have a minimum function like the "safe mode", but we have not implemented that.

2) Attack Prevention: The proposed system provides two levels of attack prevention mechanism. The first level is to prevent Normal OS from illegitimate modification. When an attacker gains full control of Normal OS to misuse the vulnerability, the attacker may overwrite the code segment of Normal OS to directly overwrite the memory. In fact, many vulnerabilities (e.g., CVE-2013-4342, CVE-2013-1969, and CVE-2008-1673) allowing a remote attacker to execute arbitrary code are reported [17]. In the case of Linux, for example, once arbitrary code is executed with an administrator privilege by an attacker, it is possible for the attacker to overwrite an arbitrary area of code segment through /dev/mem, resulting in system crash or misbehavior. Although overwriting the code segment in memory is generally difficult, it is relatively easy in the case of end-point devices since the hardware configuration is fixed. As a result, the system may go down. However, since LiSTEE^{TM} Monitor sets the access control of the memory region for the code segment of Normal OS as read-only, and its configuration can be changed only from the secure world, it is impossible for Normal OS to overwrite the code segment of Normal OS. An advantage is that the protection does not cause any side effects. Since a data segment is used to store the state of the program, Normal OS updates the content of the data segment frequently during its execution. In contrast to the data segment, since a code segment is used to store program code, it is not expected to update its content after booting the system. In particular because devices such as smart meters or concentrators are not expected to change their function after being deployed, the dynamic update function is not required. Thus, this protection mechanism can protect Normal OS from illegitimate modification without side effects. Moreover, the feature of read-only memory is vey useful for the data, whose value is only changed by LiSTEE^{TM} TA and to which Normal OS only refers. The typical application is a secure clock. In a legacy system, it is very difficult to provide a secure clock on an operating system without network connectivity or dedicated hardware if illegitimate modification of the operating system is premised. However, LiSTEE^{TM} TA can provide a local secure clock function by software. Since LiSTEE^{TM} TA is executed periodically and it knows the frequency of the execution, it is possible for LiSTEE^{TM} TA to update a counter value written in a read-only memory in a certain amout of time periodically. Because the counter value is read-only from Normal OS, Normal OS cannot revert the counter value. The second level is to protect LiSTEE^{TM} Monitor and LiSTEE^{TM} TA from illegitimate modification and suspension. Since the first level of protection is effective only for a code segment of Normal OS, an attack that overwrites a data segement cannot be prevented. Thus, there are still possibilities that control of Normal OS is gained by an attacker. Even in such cases, thanks to TZASC, since Normal OS is prohibited from overwriting the content of memory where LiSTEE^{TM} TA and LiSTEE^{TM} Monitor are allocated, illegitimate modification is prevented. Since communication interface between Normal OS and LiSTEE^{TM} TA is limited, it is impossible to compromise LiSTEE^{TM} TA by an attack. Moreover, since the interrupt configuration register is accessible only from the secure world, there is no way for Normal OS to stop the timer interrupt. Furthermore, LiSTEE^{TM} provides a mechanism to protect against shutdown attack. Since it is impossible to prevent Normal OS from executing a shutdown procedure with a priviledged instruction in the non-secure world, when a process running in the non-secure world tries to shutdown the system, LiSTEE^{TM} TA can detect it and discard the shutdown request. Since end-point devices usually keep working all the time, devices could be implemeted without having a shutdown or reboot function. However, it is necessary to have a shutdown function in some cases. For example, the system may need to reboot when updating firmware. Another example is that a service enginner may need to reboot the system when inspecting the status of the end-point devices for maintanance purposes. Although it has not been implemented, it is possible to endow LiSTEE^{TM} TA with a function to determine whether it should shutdown or not based on the status of the system. For example, when LiSTEE^{TM} TA detects an access to the memory region mapped to the registers corresponding to power management and determines that the system is under a particular status, such as a maintanance mode, it may allow executing a shutdown procedure. Similarly, when LiSTEE^{TM} TA detects the access, it sends a head-end system a message to inquire whether the shutdown request is accepted or not by using the message notification function. Based on a response to the inquiry, it can determine whether or not a shutdown procedure can be executed without interference of Normal OS.

3) System Reliability: In a legacy system, one single bug could affect the entire system, causing a critical failiure. Ideally, from a defensive viewpoint, the entire system including the operating system should be bug-free to achieve high availability. However, it is impracticable to build a complicated system without bugs. Linux 3.6.1 consists of over 15 million lines of code and many new bugs that cause critical crash are reported frequently (e.g., CVE-2013-4563, CVE-2013-4387, and CVE-2012-2127) even though it is carefully reviewed by many professionals [17]. Thus, the smaller the critical component that has to be robust within a system, the better. In the case of LiSTEE^{TM} Recovery, the critical components correspond to LiSTEE^{TM} TA and LiSTEE^{TM} Monitor. In contrast to Linux, the code size of LiSTEE^{TM} Monitor and LiSTEE^{TM} TA is relatively

small. The volume of source code for LiSTEE^{TM} Monitor is about 700 lines and its code and data size are 2.1 KB and 1.6 KB, respectively. Similarly, the volume of source code of LiSTEE^{TM} TA is about 41200 lines and its code and data size are 1.09 MB. Compared to the volume of source code of Linux, the risk of LiSTEE^{TM} Monitor and LiSTEE^{TM} TA including bugs is small.

4) Response to Failure: The Notification module in LiSTEE^{TM} TA sends a message to the head-end system just before rebooting the system. The message, which notifies that particular devices are about to reboot, is sometimes useful information for administrators. For example, if messages are sent by devices having a particular software version number, the reboot could be caused by an attack aimed at a vulnerability specific to the software. If messages are sent by devices located in one particular network, the reboot could be caused by a network worm distributed in that specific network. Although LiSTEE^{TM} Recovery cannot prevent an attack in advance, the notification feature can help the administrator investigate the reason for the failure during or after the incident. For example, it is impossible for LiSTEE^{TM} Recovery to prevent an attacker from compromising Normal OS and causing reboot frequently. However, the administrator can notice that frequent reboot occurs to the device through network since Notification module sends a message each time when rebooting. The attackers may try to block sending of the message to circumvent the notification. However, Normal OS cannot interfere with the Notification module sending a message to the head-end server since the Notification module is executed inside LiSTEE^{TM} TA. Moreover, since LiSTEE^{TM} TA is processed in an environment isolated from Normal OS, security processes, such as encrypting a message, are easy to implement in LiSTEE^{TM} TA. Therefore, once an encryption key and an encryption process are implemented in LiSTEE^{TM}, it is possible to keep them secret from Normal OS. In the next step, it is possible to include a firmware update feature to implement functions receiving data from the head-end system and writing the data into the file system to extend the function of the Notification module. In combination with the "safe mode" described above, this function is effective against a continuous attack that occurs just after the system recovers.

B. Performance Analysis

As well as the implementation environment, we used Motherboard Express uATX that contains the ARM Cortex-A9x4 processor running at 400 MHz as an experimental environment. Level 1 instruction cache, level 1 data cache, and level 2 cache are 32 KB, 32 KB, and 512 KB, respectively. It contains 1 GB DRAM as the main memory and we assigned the same memory map as that described in Section V.

First, we measured the execution time of LiSTEE^{TM} TA during execution of Normal OS; to be precise, the time period from the beginning of the hardware interrupt handler in LiSTEE^{TM} Monitor through to the execution of the SMC instruction. Without calling the Notification module, the average time is 1.7 microseconds over 10,000 trials. However, if the Notification module is called, the average time is 4.1 milliseconds over 10,000 trials. Note that the Notification module is called when rebooting the system, which rarely occurs. Thus, this performance overhead poses no problem.

Next, we measured the performance degradation of Normal OS. Since the execution of Normal OS is suspended during execution of LiSTEE^{TM} TA, the performance of Normal OS degrades in any case. The total of Normal OS suspension time depends on the frequency of calling LiSTEE^{TM} TA. There is a tradeoff between the performance degradation of Normal OS and the delay in detecting the crash of Normal OS. When the frequency is increased, the performance degradation of Normal OS is also increased. On the other hand, when the frequency is decreased, the delay for detecting the crash of Normal OS becomes larger. Since a general application is assumed to be executed on Normal OS, we used dhrystone as a benchmark program to measure the performance degradation [18].

Fig. 7 shows the result of the experiment. The bar graph

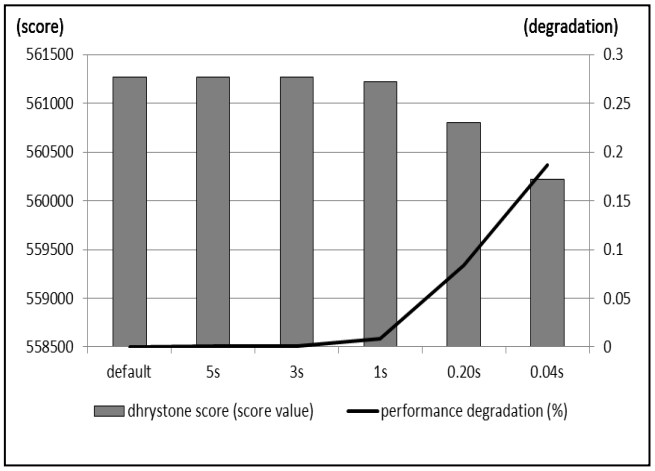


Figure 7. Result of performance degradation.

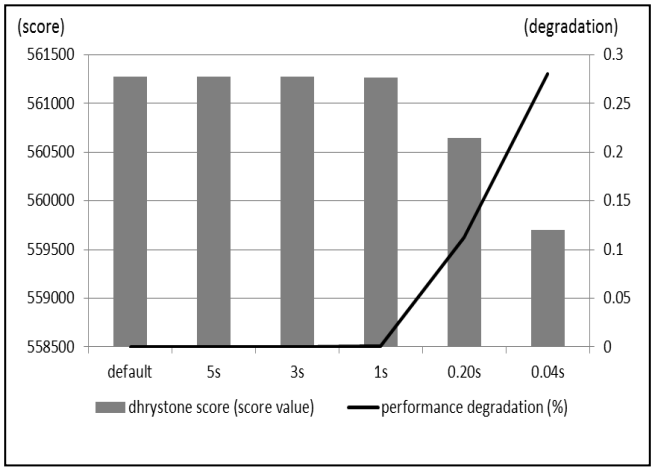


Figure 8. Result of performance degradation with message transmission.

shows the dhrystone score and the line graph shows the performance degradation. The higher the score, the better the performance is. Each bar shows the timer interval of calling LiSTEE^{TM} TA and its value is default (never called), 5 seconds, 3 seconds, 1 second, 0.2 seconds and 0.04 seconds respectively. When the timer interval was set to 5 seconds, the performance degradation was suppressed within 0.001 %. Even if the interval was set to 0.04 seconds, the performance degradation was less than 0.2 %. The result shows that although there is a tradeoff between performance degradation of Normal OS and detection rate logically, the performance degradation can be ignored in practice even if the frequency of calling LiSTEE^{TM} TA is increased. Fig. 8 shows another result of the experiment. In the case of Fig. 7, it is assumed that the Notification module sends a head-end system a message only when Normal OS stops working and the system is rebooting. Therefore, the result does not include processing time of the Notification module. On the other hand, Fig. 8 assumes that the Notification module sends a head-end system a 32 byte message whenever LiSTEE^{TM} TA is executed even if Normal OS is working correctly. This experiment assumes that Notification module sends the head-end system a message periodically even if Normal OS keeps working so that an administrator can monitor the status of each device. Although the result of the experiment shows that the performance slightly degrades compared with the experiment without message transmission, it can still be ignored in practice. Note that the score was better for the experiment with message transmission than for the experiment without message transmission when the interval was set to 5 seconds, 3 seconds and 1 second. When the timer interval is long, the execution times of LiSTEE^{TM} TA and LiSTEE^{TM} Monitor are negligible compared with the execution time of Normal OS since the task is too small to measure accurately. Thus, this can be regarded as an error.

C. Cost Analysis

1) Development Cost: LiSTEE^{TM} Recovery does not require any modification to Linux in order to run it as Normal OS on LiSTEE^{TM} Monitor. Thus, in terms of application developer's cost, since developers can reuse all existing programs including libraries, middleware, and applications running on Linux, no additional develpment cost is necessary. In terms of device developper's cost, configuration, such as network address setting of Notification module, and memory address setting and security permission setting of TZASC is necessary to integrate LiSTEE^{TM} Recovery into a device. In addition to the development cost, verification cost in order to check that the configuration is correct is necessary. For embedded devices in Smart Grid, there are cases where the performance requirement is specified. For example, in the case of a smart meter, it is reported that an acceptable delay in responding to a management server is in the range of 50 ms to 300 ms under a specific condition [19]. As described in the performance analysis, since performance degradation is insignificant when introducing our proposed method, the cases requiring performance tuning are limited. Therefore, the development cost can be controlled.

2) Production Cost: LiSTEE^{TM} Recovery is software-based technology and no additional hardware except a TrustZone-capable ARM processor and an address space controller is required. TrustZone-capable processors are widely available. In fact, all ARM Cortex A series processors support TrustZone. Therefore, the additional cost is mitigated. As a result, development cost per device can be minimized.

3) Maintenance Cost: It is assumed that a tremendous number of devices will be deployed in the field for smart grids. In the case of a cyber-attack, since many devices could be a target of the attack and the attack could be done in a very short period of time through the network, it is impracticable in terms of both cost and time for field service engineers to physically visit each site and reboot them. The autorecovery feature of LiSTEE^{TM} Recovery mitigates this problem. Moreover, the report is sent to the head-end system once the device reboots. This function contributes to reduction of the cost of troubleshooting. Thus, LiSTEE^{TM} Recovery provides an opportunity to reduce maintainance cost compared with legacy systems.

VII. Related work

To recover from an operating system failure, various approaches have been proposed.

The simplest approach is that of including the recovery mechanism within the operating system. One method is to use Non-maskable Interrupt (NMI) as a watchdog timer [20]. NMI is a processor interrupt that cannot be ignored. When NMI is generated, the NMI handler implemented within the operating system is called regardless of the status of the operating system. Since it is not necessary to save and restore registers to execute a process implemented in NMI handler, performance overhead is mitigated. Thus, NMI can be used as a surveillance and recovery process to implement the NMI handler so that it detects whether the operating system hangs or not. In [21], Dolev et al. propose self-stabilizing operating system by utilizing NMI. Although NMI is easy to use as a watchdog timer because it has already been implemented in Linux, it is vulnerable because the NMI handler could be invalidated to overwrite the code segment of the operating system. Furthermore, since implementation of a rich application in an interrupt handler, such as a network communication function or a data encryption function, is not anticipated, it is difficult to realize the notification function.

Another approach to recover from the failure is to check the status of the operating system from outside using virtualization technology. It is easy to realize an isolation environment by utilizing virtualization technology. Karfinkel developed the trusted virtual machine monitor (TVMM), on which a general-purpose platform and a special-purpose platform executing security-sensitive processes run separately and concurrently [22]. The libvirt project develops a virtualization abstraction layer including a virtual hardware watchdog device [23]. To cooperate with the watchdog

daemon installed in a guest OS, a virtual machine monitor can notice that the daemon is no longer working when periodically trying to communicate with it. Although virtualization technology is widely deployed in PC-based systems, it is difficult to implement it in embedded devices as fewer hardware devices support it. Moreover, since the volume of source code for a virtual machine monitor (VMM) tends to become large, the risk of VMM including bugs also becomes large. To overcome the restriction, Kanda developed SPUMONE, which a lightweight virtual machine monitor designed to work on embedded processors [24]. It provides a function to reboot the guest OS. However, SPUMONE does not provide a memory protection mechanism between the virtual machine monitor and the guest OS (Normal OS). Thus, it is vulnerable to an attack on the virtual machine monitor from the guest OS.

To make a secure environment by utilizing TrustZone, various systems have been proposed.

In [25], Yan-ling et al. propose a secure embedded system environment with multi policy access control mechanism and a secure reinforcement method based on TrustZone. It assumes various applications and services runs on it. In [26], Sangorrin et al. propose a software architecture on which real-time operating system and a general-purpose operating system are executed concurrently on a single ARM processor with low overhead and reliability by utilizing TrustZone. Baseline common functions described in Section IV basically uses the same technique in the existing approaches. Our contribution is clarifying a total architecture and functions which must work in a secure environment with a full implementation to enable end-point devices automatically to recover from an error status in a Smart Grid.

VIII. CONCLUSION AND FUTURE WORK

LiSTEE^{TM} Recovery works effectively to resist critical bugs or attacks including zero-day attacks, that could potentially cause the system to crash, in order to keep availability of end-point devices. The performance evaluation has been presented to show that the degradation of the existing system is sufficiently small. Considering commercialization, we have shown that the development cost and production cost can be minimized. Moreover, LiSTEE^{TM} Recovery can save maintenance cost.

Future work includes resistance to sophisticated attacks. In one possible attack, an attacker illegitimately modifies the shared memory area to fake as if Normal OS works correctly while almost all Normal OS functions actually stop. As a result, LiSTEE^{TM} TA misunderstands that Normal OS works correctly. One approach to solve this attack is to implement LiSTEE^{TM} TA so that it itself checks the status of Normal OS without the support of an application program running on Normal OS. For example, whenever Normal OS is running, it must update a certain data area, such as page tables or process tables. Therefore, in monitoring the data area, LiSTEE^{TM} TA can determine whether Normal OS is working or has crashed. An advantage of LiSTEE^{TM} is that it is impossible for an attacker to reverse-engineer and to tamper with an algorithm of LiSTEE^{TM} TA from Normal OS because of the memory protection mechanism. Thus, an attacker cannot know how to compromise Normal OS in order to produce misleading information. We have not implemented this though. Another possible attack involves damaging the file system locating Normal OS. Network boot can be a solution where LiSTEE^{TM} TA downloads a small rescue program from the head-end system when booting fails.

REFERENCES

[1] Isozaki, H., Kanai, J., Sasaki, S. and Sano. S., “Keeping High Availability of Connected End-point Devices in Smart Grid,” In Proc. Fourth International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies, Apr. 2014, pp. 73-80.

[2] C4 Security. “The Dark Side of the Smart Grid,” [Online]. Available: http://www.c4-security.com/The Dark Side of the Smart Grid – Smart Meters (in)Security.pdf [Accessed 20 Nov. 2014]

[3] Forsberg, D., Ohba, Y., Patil, B., Tschofenig, H., and Yegin, A., “Protocol for carrying authentication for network access,” IETF RFC 5191 [Online]. Available: http://tools.ietf.org/html/rfc5191 [Accessed 20 Nov. 2014]

[4] Zhao, F., Hanatani, Y., Komano, Y., Smyth, B., Ito, S., and Kambayashi, T., “Secure authenticated key exchange with revocation for smart grid,” The third IEEE PES Conference on Innovative Smart Grid Technologies (ISGT 2012), IEEE Power & Energy Society (PES), Jan. 2012, pp. 1-8.

[5] Wang, W. and Lu, Z., “Cyber security in the Smart Grid: Survey and challenges,” Computer Networks, Vol. 57, Issue 5, Apr. 2013, pp. 1344-1371.

[6] Khurana, H., Hadley, M., Ning, L., and Frincke, D.A., "Smart-grid security issues," IEEE Security & Privacy, Vol. 8, Issue 1, Jan-Feb. 2010, pp. 81-85.

[7] Li. K, "Towards Security Vulnerability Detection by Source Code Model Checking," Software Testing, Verification, and Validation Workshops (ICSTW), 2010 Third International Conference on, Apr. 2010, pp. 381-387.

[8] Rinard, M., Cadar, C., Dumitran, D., Roy, D. M., and Leu, T., "A Dynamic Technique for Eliminating Buffer Overflow Vulnerabilities (and Other Memory Errors)," Computer Security Applications Conference, 2004. 20th Annual, Dec. 2004, pp. 82-90.

[9] Huang, Z and Harris, I.G., “Return-oriented vulnerabilities in ARM executables,” IEEE 2012 Conference on Technology for Homeland Security, Nov. 2012, pp. 1-6.

[10] De Craemer, K., and Deconinck, G., "Analysis of state-of-the-art Smart Metering communication standards,” in Young Researchers Symposium (YRS), 2010. [Online]. Available: https://lirias.kuleuven.be/bitstream/123456789/265822/1/SmartMeteringCommStandards.pdf [Accessed 20 Nov. 2014]

[11] Wang, W., Xu, Y., and Khanna, M., "A survey on the communication architectures in smart grid," Computer Networks, Vol. 55, Issue 15, Oct. 2011, pp. 3604-3629.

[12] Liotta, A., Geelen, D., van Kempen, G., and van Hoogstraten, F., “A survey on networks for smart-metering systems,” International Journal of Pervasive Computing and Communications, Vol. 8, No.1, 2012, pp. 23-52.

[13] Varghese, S. M., and Jacob, K. P., "Anomaly Detection Using System Call Sequence Sets," Journal of Software, Vol. 2, No. 6, Dec. 2007, pp. 14-21.

[14] Pont, M. and R. Ong., “Using watchdog timers to improve the reliability of single-processor embedded systems: Seven new patterns and a case study,” In Proc. First Nordic Conf. on Pattern Languages of Programs, Sept. 2002, pp. 159-200.

[15] ARM. “ARM Security Technology,” [Online]. Available: http://infocenter.arm.com/help/topic/com.arm.doc.prd29-

genc-009492c/PRD29-GENC-009492C_trustzone_security_whitepaper.pdf [Accessed 20 Nov. 2014]

[16] Alves, T. and Felton, D., "TrustZone: Integrated Hardware and Software Security," Information Quarterly, Vol. 3, No. 4, 2004, pp. 18-24.

[17] MITRE. "Common vulnerabilities and exposures," [Online]. Available: http://cve.mitre.org [Accessed 20 Nov. 2014]

[18] ARM. "Dhrystone Benchmarking for ARM Cortex Processors," [Online]. Available: http://infocenter.arm.com/help/topic/com.arm.doc.dai0273a/DAI0273A_dhrystone_benchmarking.pdf [Accessed 20 Nov. 2014]

[19] Miyashita, M. and Ohtani, T., "Transmission Characteristics Evaluation of Demand-side Communication -Evaluation of Response Time Using International Standard Protocol for Meter Reading and Wireless LAN-," CRIEPI Research Report, Jun. 2011 (in Japanese).

[20] Kleen, A., "Machine check handling on linux," Technical report, SUSE Labs, Aug. 2004 [Online]. Available: http://halobates.de/mce.pdf [Accessed 20 Nov. 2014]

[21] Dolev, S. and Yagel, R., "Towards Self-Stabilizing Operating Systems," IEEE Transaction on Software Engineering, Vol. 34, No. 4, Jul/Aug. 2008, pp. 564-576.

[22] Garfinkel, T., Pfaff, B., Chow, J., Rosenblum, M., and Boneh, D., "Terra: A virtual machine-based platform for trusted computing," In Proc. Symposium on Operating System Principles, Oct. 2003, pp. 193-206.

[23] "libvirt - the virtualization API.," [Online]. Available: http://libvirt.org [Accessed 20 Nov. 2014]

[24] Kanda, W., Yumura, Y., Kinebuchi, Y., Makijima, K., and Nakajima, T., "SPUMONE: Lightweight CPU Virtualization Layer for Embedded Systems," In Proc. Embedded and Ubiquitous Computing, Dec. 2008, pp. 144-151.

[25] Yan-ling, Z., Wei, P., "Design and Implementation of Secure Embedded Systems Based on Trustzone," In Proc. International Conference on Embedded Software and Systems, Jul. 2008, pp. 136-141.

[26] Sangorrin, D., Honda, S. and Takada, H., "Dual Operating System Architecture for Real-Time Embedded Systems," In Proc. 6th International Workshop on Operating Systems Platforms for Embedded Real-Time Applications, Jul. 2010, pp. 6-15.

Multi-Protocol Transport Layer QoS: An Emulation Based Performance Analysis for the Internet of Things

James Wilcox
Center for Communications Research
University of Bristol, UK
james.wilcox@bristol.ac.uk

Dritan Kaleshi
Center for Communications Research
University of Bristol, UK
dritan.kaleshi@bristol.ac.uk

Mahesh Sooriyabandara
Telecommunication Research Laboratory
Toshiba Research Europe Limited, UK
Mahesh@toshiba-trel.com

***Abstract*—Application specific interaction models will be required to support efficient communications between distributed applications with disparate network requirements within the Internet of Things. This paper shows that intelligently selected transport protocols are able to provide increased efficiency of network resource usage under specific network conditions. Real-time adaptive selection of transport protocols makes it possible to achieve a distributed embedded system with heterogeneous actors that can react to both application-specified Quality of Service (QoS) requirements and varying network conditions. vNET, a custom, virtualisation based, distributed network emulation test bed will be presented and validated using an MQTT performance analysis before using it to validate the premise of multi-protocol transport layer QoS.**

Keywords–Quality of Service, Internet of Things, MQTT, Adaptive Transport Layer, Network Emulation, Middleware, Smart Grid

I. INTRODUCTION

This paper is an extended piece based on [1]. It provides additional content on the virtualisation based network emulation test bed developed for this work.

Traditionally networked applications are designed with a pre-selected transport layer protocol. Optimisations for a specific application are done at the application layer and all messages are transported using the same protocol, either TCP (Transmission Control Protocol) [2], UDP (User Datagram Protocol), or with an overlay transport protocol such as RTP (Real Time Protocol) [3]; Figure 1 represents this paradigm. This is typically fixed at application development; however, there is no fundamental requirement for this to be the general rule. Whilst networked applications need to exchange information, there is no reason why application layer code should be concerned with how that information is transported. There are a multitude of existing, mature transport layer protocols available each designed to tackle specific network problems [4]. Utilising these many protocols, a single application could leverage the advantages of each protocol individually at the appropriate time given an environment with dynamic network conditions and application requirements. Acknowledging these points raises the challenge of defining a generic framework that allows for run-time selection of transport protocols to dynamically match specific application requirements and, specifically, message patterns used by the application. If the low level network interactions enforced by a specific transport protocol and higher level architectural messaging pattern are completely decoupled from the application then dynamically modifying the combination can be used as standard. If certain transport protocols and messaging pattern combinations are able to provide higher performance in terms of bandwidth, latency and reliability in certain network environments than others can do, then by supporting adaptive selection of these combinations it becomes possible to have a distributed real-time embedded (DRE) system with heterogeneous actors that can react to both dynamic application QoS requirements and network conditions. This model is shown in Figure 2. The middleware system required for managing the selection of the large numbers of transport protocols referenced in Figure 2 has been developed and presented in DIRECTOR: A Distributed Communication Transport Manager for the Smart Grid [5].

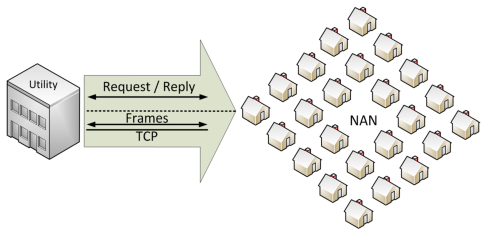


Figure 1. Traditional: Applications are supported by a single interaction model (in this case, TCP Request / Reply). The utility represents systems providing the back end infrastructure.

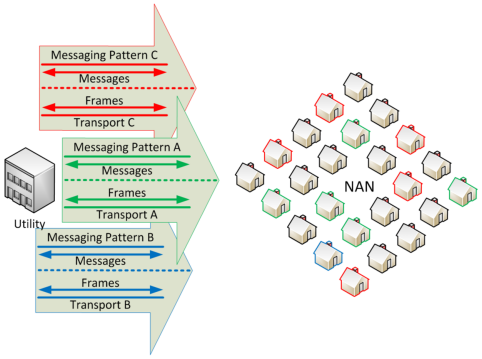


Figure 2. Proposed: Applications using multiple run time optimised interaction models (managed using middleware) instead of only TCP Request / Reply.

An ”interaction model” as used throughout this work describes the virtual patterns of packets, datagrams or frames (transport protocol and messaging pattern combination) that

constitute a data communication session between at least two networked devices. A specific interaction model will communicate over compatible interfaces. For example there is a clear difference in interaction model between communicating with unicast and communicating with multicast transports. Figure 3 further demonstrates how multi-protocol QoS could be used to exploit the under utilised optimisation opportunities using a more specific example. Both approaches shown achieve the same overall goal - the three receiving nodes obtain the sending nodes data. However, in this scenario, multicast offers clear bandwidth efficiency increases compared to unicast. Therefore, an intelligent system should be able to automatically make this determination to improve performance in a selected metric. This premise extends much further than unicast, multicast and bandwidth efficiency however. There are numerous approaches to getting packet X to destination(s) Y. Each interaction model has its own set of attributes which affect the performance of both the application and the underlying network in different ways. The attributes can be classed and characterised to provide groups of interaction models that provide specific advantages in certain communication scenarios. So while it is possible to support many disparate IoT applications with a fixed interaction model, typically tied in with a single transport protocol, it would be more efficient in terms of network resources to be able to provide tailored interaction models on a per-application basis. These tailored interaction models would be designed to meet specific Quality of Service (QoS) levels utilising network resources more efficiently than the case when only a fixed interaction model approach, often set at system design stage.

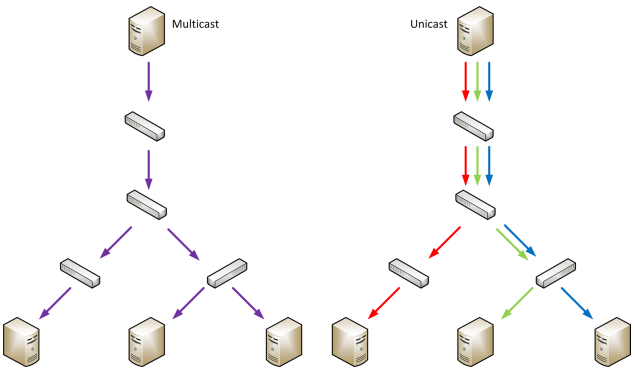


Figure 3. Interaction model differences for disseminating the same data using Multicast vs. Unicast

This paper shows experimental results which demonstrate that specific interaction models will provide performance gains over a pre-defined communication transport architecture and that certain combinations provide useful gains over other potentially viable options. The potential of generating a large scale mapping of transport combinations and application requirements will be explored. The main contributions of this paper are:

1) An analysis of performance of a distributed system for different interaction models by running application scenarios using a fixed, realistic, network topology.
2) Results that show that the performance of the Distributed Real-time Embedded (DRE) applications varies significantly for different interaction models, suggesting that this then can be used to optimise the performance of the individual transactions that make up the application network traffic.
3) Presentation of vNET, a virtualisation based, distributed network emulator.

The paper is organized into the following sections: section II provides the related material and further motivation for this work. Section III presents the experimental emulation test bed setup and the viable communication interaction models. Section IV presents the experimental parameters and the results. Section V presents a potential middleware solution that can capitalise on these findings. Section VI presents the conclusions from the experiments and the direction of future work.

A. Related Work and Motivation

Environments targeted by this work have the follow characteristics:

1) They are distributed and built from a large number of heterogeneous embedded devices, running a number of different applications.
2) They are typically loosely-coupled.
3) The majority of the actors are communication network-constrained rather than resource-constrained.
4) Each device is expected to run many different applications with varying network requirements.

One example of such a system is the SG, and in particular, the subset of applications that intend to use consumer / demand side equipment and systems to achieve grid specific goals such as load shedding or load shifting or in more general terms, Demand Side Management (DSM). Section A introduces the SG and its edge applications. Several related works exist which support with the premise of providing DRE software applications, such as those that will operate in a SG, with flexible communication choices in order to either achieve better network resource allocations or meet specific communication requirements. These overviews are presented in sections B and C.

1) The Smart Grid and Demand Side Management (DSM): The SG can be seen as a large scale distributed IoT system, with a large component being embedded sensor-actuator networks to support distribution power network monitoring and control and DSM interactions. DSM focuses around the control of demand side loads in the electricity distribution network in order to manipulate network conditions [6]. DSM breaks down into a number of related but still significantly different enough sub-applications to warrant different communications approaches. DSM can be broken down into two major sub categories, Demand Control (DC) [7] and Demand Response (DR) [8]. DC is defined as DSM programs that have centralized direct control over consumer loads [7]; DR is defined as DSM programs that use indirect methods (typically pricing) to affect changes [9]. Each approach requires a different communications paradigm in order to utilise network resources efficiently and operate optimally. The above presents an ideal system for this work. It presents the rare opportunity to take a completely different approach to facilitating machine-to-machine communications in a DRE environment. The SG will eventually call for millions of networked geographically-distributed embedded

devices to be deployed into the demand side of the power distribution grid. These devices are either designed to utilise existing networks such as domestic broadband or cellular networks and coexist with the existing traffic, or to utilise purpose built resource constrained networks such as various forms of wireless mesh or power line communications [10]. Both approaches result in strict network resource constraints for the applications. These constraints further increase the impact run-time transport level adaptive QoS will have in such environments. The main argument against dynamically matching interaction models to application requirements and real time network conditions has been one of complexity. With sensor networks, the SG and the Internet of Things (IoT) in general becoming more prevalent, the environment is changing and these arguments are no longer valid. Our proposition is that the performance gain introduced by dynamically matching interaction models outweighs the required increase in complexity of the architecture and embedded hardware. DSM can be shown to be a good example of how tailored communication paradigms could be beneficial in the SG and similar environments.

2) Transport Mechanisms : The concept of adaptive transport layer services for resource-constrained environments is well explored; however, the approach taken usually considers the transport protocol to be used already pre-selected at development stage, and only considers adaptations above the transport layer. They do not propose to provide application optimisations at the lower transport level. However, applications, regardless of the type of resource-constrained environment, can benefit from tailored communication service at the transport layer. Mutlu et al. [11] presents a middleware solution for performing transport level QoS focused on Bluetooth application profiles and uses CORBA (Common Object Request Broker Architecture) [12] to facilitate the middleware. While the scope is clearly limited, and transport protocol choices are not part of the QoS mechanism, the motivation is similar. Furthermore, it can be shown that different communication protocols have inherently different QoS characteristics and that using targeted protocols with specific applications can improve performance with a number of chosen metrics. Weishan et al. [13] recognise this and provide experimental results related to protocol switching overhead and also implement the system using a middleware solution. They conclude that protocol switching overhead is minimal with their chosen transport protocols and that protocol switching is beneficial to DRE environments.

3) QoS Architectures for DRE systems: The works highlighted here are attempting to improve or maintain DRE application performance in sub-optimal or resource-constrained networks by utilising real-time adaptive QoS management mechanisms. [14], [15], [16] focus on a single interaction model and attempt to provide adaptive QoS within these confines. It demonstrates that additional QoS optimisation opportunities are available if the scope of the system includes controlling lower level attributes such as interaction models in conjunction with the adaptive QoS mechanisms. For example Wenjie et al. [15] propose a QoS adaptive framework for Publish-Subscribe Service called QoS Adaptive Publish-Subscribe (QAPS). They define several QoS policies and focus on fault tolerance and dependability of services. Schantz et al. [17] present a distributed, real-time embedded system capable of adaptive QoS. They describe in detail several methods of implementing end-to-end adaptive QoS mechanisms and explain how the work gives DRE applications more precise control over how their end-to-end resource allocations are managed. These proposed adaptive QoS mechanisms all address the same problem as this paper, but these implementations are limited to the application layer instead of considering a multi-protocol transport layer to access additional optimisation opportunities. Zieba et al. [16] develop the concept of quality-constrained routing in publish / subscribe messaging architectures. They develop a system which integrates application quality requirements into the message routing architecture in order to better support dealing with varying network conditions such as dynamic network topologies and link characteristics. The idea of integrating the dynamic application requirements into the communication paradigm provides a critical distinction from the others and further reinforces the need for verified optimised communication paradigms in order to meet these dynamic requirements.

II. vNET: A Virtualised Network Emulation Test-bed for the Embedded Internet of Things

Developing and evaluating network solutions for IoT environments, especially those near deployment, requires a method of allowing the participating entities and critical hardware components to interact in a controllable, scaled way. A real world trial would be ideal as it would produce highly detailed and accurate results but the cost of this approach is often prohibitive. Another solution is to run simulations which are cheap, especially if there is no real-time requirement, but only provide results as accurate as the models used and the assumptions made. There are many network simulation tools available such as Qualnet and NS2/3 [18], [19] that already have tried and tested networking models for a variety of scenarios. However, for systems that can be described as *complex networks of discrete components* like multi-application IoT environments, simulation does not naturally lend itself. Providing models for each disparate networked entity is time consuming and furthermore if the goal is to evaluate how software or devices that utilise custom interaction models perform in a distributed network, simulating these entities would often require re-implementing the network code so that it is compatible with the simulator. This is inefficient and difficult especially if the system is required to react to unscripted network events.

Network *emulation* provides a viable and valuable alternative in these cases. With emulation, the network and the devices that a system consists of are completely represented and inherently provide the same interfaces and functionality the real world system would. This allows real development code to be directly evaluated in an easily controllable and scalable manner without resorting to a physical deployment. The emulation test-bed can be seen as a condensed version of a real life system with all the varying levels of complexity a real world system would have, from the standard open source software running on each node down to the physical layer of the network. In addition to this, the ability to pass real hardware to specific nodes within the network allows OS driver interactions and hardware choices to be evaluated potentially revealing incompatibilities and areas for optimisation before moving onto scaled real-world trials. Until relatively recently,

full network emulation on a useful scale for evaluating DRE environments has been difficult to achieve, mainly due to computing resource constraints. Recent advances in virtualisation technology can be used to address this problem. Furthermore, as IoT devices are typically resource constrained and relatively low performance, large numbers can be fully emulated with modest, and therefore inexpensive, host hardware. This makes an emulation test-bed a valuable and more suited tool in evaluating embedded distributed networks compared to simulation. Furthermore, real network hardware is incorporated into the test-bed as a proof-of-concept for hardware-in-the-loop (HIL) emulation based testing which provides the ability to evaluate real hardware and driver choices for possible incompatibilities.

This section presents:

1) A virtualisation based network emulation test bed that allows unmodified code and real hardware / driver interactions to be evaluated in an extensible, fully controllable distributed, software defined, networked environment.
2) A series of experimental scenarios based on the IBM MQTT performance analysis [1]. These experiments will focus on analysing CPU and bandwidth consumption of the various QoS levels MQTT provides and will demonstrate the functionality and capabilities of the test-bed.

A. Network Emulation Related Work

Basic requirement set for a viable distributed network, embedded device test-bed:

- Complete flexibility in terms of evaluating software in both user and kernel space.
- Full integration with physical networks.
- Ability to evaluate real hardware and driver choices.
- Scalable in terms of node numbers.
- Low emulation overhead.

Several candidates were identified which met some of these requirements. The rest of this sections evaluates these potential choices.

Network analysis through the use of emulation and virtualisation is not a new area [20], [21], [22], [23]. There is a large amount of work that uses network emulation due to the benefits it provides over purely simulation based analysis.

The Common Open Research Emulator (CORE) used by [23] provides an ideal example of the capabilities of existing emulation test-beds. CORE is the closest comparable technology to the emulation test bed described in this work. However, there are 2 areas where CORE lacks features that allow access to previously unexplored analysis opportunities:

1) CORE does not support passing through physical hardware such as NICs, storage controllers and other application specific hardware.
2) The LXC (Linux Containers) used by CORE *must* use the same host system kernel and so each emulated node cannot use a custom kernel or kernel level modifications and this limits flexibility.

Mininet [24] is another network emulator that uses LXC. It is provides the ability to evaluate large networks using consumer grade PC hardware. Mininet explicitly targets Software Defined Network (SDN) environments and allows vendor independent OpenFlow interface compatible controllers to be experimented with. While Mininet, like CORE, are extremely useful network emulation tools, they are still limited by LXC. While LXC allows greater numbers of nodes to be emulated, the lack of complete isolation and ability to implemented kernel level or individual network stack modifications is clearly limiting.

Cooja [25] is a Java based contiki [26] mote simulator. Contiki is an OS for small embedded sensor platforms and Cooja was developed in order to evaluate interactions between small numbers of them without resorting to physical trials. Cooja is able to simulate at multiple layers including, the network layer, the OS layer and the machine instruction set layer if required. While Cooja is ideally suited for testing small ($<$100) node mote networks, its usefulness outside such network environments is limited.

TABLE I. Related Network Evaluation Tools [27] [28]

	Key Tech	Node Limits	Network Limits	Target Scens	Comments / Use cases
CORE	LXC, NS	No hard limit	Not Real Time	Non specific	Limited interactions with physical networks
Cooja	Java, NS	$<$100 nodes	Mote Interfaces	Contiki wireless mesh	Limited to evaluating motes.
Mininet	LXC	4096 per host	4Gbit/s with 4GHz CPU core	SDN OpenFlow	Large scale topology evaluation, SDN algorithms, Limited interactions with physical networks
vNET	ESXi, VMs	512 per host	9.5Gbit/s vmxnet3	Near deployment testing	Embedded Hardware Evaluation. Efficient interactions with physical networks. Software flexibility.

Expanding on these points:

Using a virtualisation technology that can take advantage of AMD-V's IOMMU or Intel's VT-d virtualisation technologies would allow real hardware devices to be passed to a virtual machine. The virtual machine kernel recognises the hardware as it would in a real system and loads the standard drivers for its operation. This allows the intricacies of real driver / hardware interactions to be evaluated. For example, a previously emulated key backbone connection could be seamlessly brought out into the real world using physical connections and network interfaces so that hardware chip sets and driver optimisations could be tested under varying load levels before resorting to comparatively expensive and time consuming real world scaled trials. This physical hardware can quickly be assigned to any hosted virtual machine and used natively. This is especially useful for systems that are near deployment.

LXC does not provide *virtual machines*, it provides a *virtual environment*. Therefore, the major problem with using LXC and similar container based virtualisation technologies, is the lack of complete isolation between virtual machines. Each LXC container shares the same kernel with one another and also the host itself. This means all software must be compiled for the same CPU architecture and kernel modifications between nodes are not possible. Emulating a network of disparate devices with this technology is therefore more difficult and restrictive.

For these reasons a new and custom fully virtualised network emulator was required.

B. Test-bed Components

The following section describes the various software tools and components that make up the virtualised emulation test bed. The following components are discussed:

1) *Virtualisation hypervisor*. Provides the virtual distributed environment and resource usage monitoring.
2) *WANem*. Provides control over the behaviour of specific individual communication links and network segments
3) *TCPDump and WireShark*. Provides data capture and offline analysis.
4) A set of custom Windows Power Shell VMware vSphere CLI scripts for deploying and manipulating custom virtual machines and their allocated resources quickly and efficiently.

C. Hypervisor Choice

There are a number of commercial and open source virtualisation technologies available. Specifically a native or 'bare metal' hypervisor was desired over a 'hosted' hypervisor in order for the computational virtualisation overheads to be as small as possible.

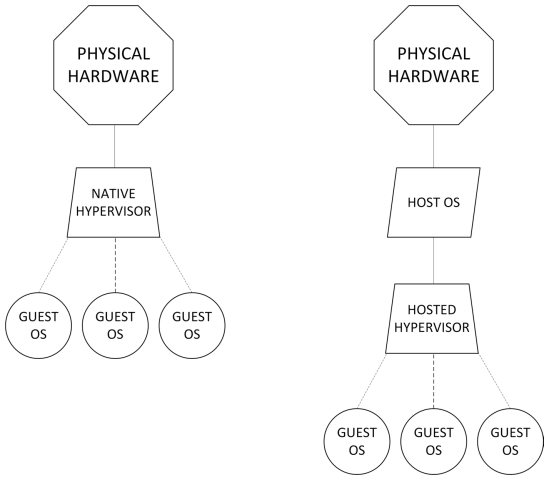


Figure 4. Native (bare metal) vs. hosted hypervisors.

Figure 4 shows the key difference between the choice of native and hosted hypervisors. A hosted hypervisor requires the support of an underlying general purpose operating system (OS), or *host OS*. This host OS requires computational resources such as CPU time, storage I/O and Memory I/O in order to operate and this means these resources are not available to the hosted hypervisor guest OS. A bare metal or native hypervisor does not require this support and therefore more resources are available to the guest OS. The requirement of a bare metal hypervisor limits the technology choices to:

1) VMWare ESXi
2) Xen Hypervisor
3) Microsoft Hyper-V

Each of these technologies has the same basic functionality and could have been used to implement the emulation test-bed. ESXi is largely considered the most mature product within enterprise and provides a robust VM management infrastructure that would make implementing and operating the test-bed easier. For these reasons ESXi was chosen as the emulation test beds native hypervisor.

D. Virtualised Distributed Environment - ESXi

Enterprise-class virtualisation provides a convenient approach to large scale cheap and efficient emulation. The test-bed described in this paper uses the VMware ESXi 5.1 bare metal hypervisor virtualisation technology [29]. ESXi allows the complete virtualisation of x86-64 based computers on a large scale. The proliferation of enterprise class technology such as Intels VT-x and AMDs AMD-V into consumer PC hardware provides dedicated on die hardware for accelerating virtualisation operations. This significantly improves the performance of virtualisation over previous generations of hardware. While typically this virtualisation technology is used to consolidate services onto a single set of server grade hardware it also has the potential to be used in unconventional ways. In this case the technology has been used to host 350 micro guests. Each of these guest nodes can be allocated a fixed amount of computing resources and / or physical hosted hardware in order to emulate a specific real world resource constrained device. Figure 5 shows an overview of the test environment. Each network segment represents an isolated network. Each network segment has an network bridge which connects it to neighbouring segments. Traffic is controlled here to emulate various network types and conditions. The system is managed by an external desktop computer which runs the VMware vSphere client and executes the custom CLI scripts detailed in section II-G. WAN access through Network Address Translation (NAT), and DHCP and DNS services are optionally provided to experimental nodes using a dedicated VM running the open source pfsense [30] 2.1 routing OS. The diagram shown in Figure 5 shows a network topology for emulating a Open Automated Demand Response Real Time Pricing usecase (OpenADR RTP) [31] but the topology is easily reconfigured through the vSphere software.

The physical hardware of the host consists of a 3.12GHz 8 core AMD FX-8120, 32GB DDR3 memory and 960GB of Solid State Drives (SSD). This hardware provides enough computing performance to support 350 fully implemented nodes. This hardware is now several generations old. Commercial offerings from both AMD and Intel have increased both Instructions Per Clock and base clock speeds of their processors. This means that for the same cost more nodes can now be supported.

E. Network Connectivity Emulation WANem

Network connectivity is emulated through the use of virtualised switches, traffic shaping and virtualised software network bridges. Complex network topologies can be emulated by introducing these virtual components at specific points in the virtualised environment. Traffic shaping is provided by WANem [32] a software WAN emulator. It provides the ability to manipulate many common network characteristics including bandwidth limitation, latency, packet loss and random network disconnections. Non ideal networks can be emulated by setting various levels of packet loss and packet corruption and this allows a system's behaviour to be evaluated under sub optimal conditions. Multiple, dedicated TCPDump traffic sinks are

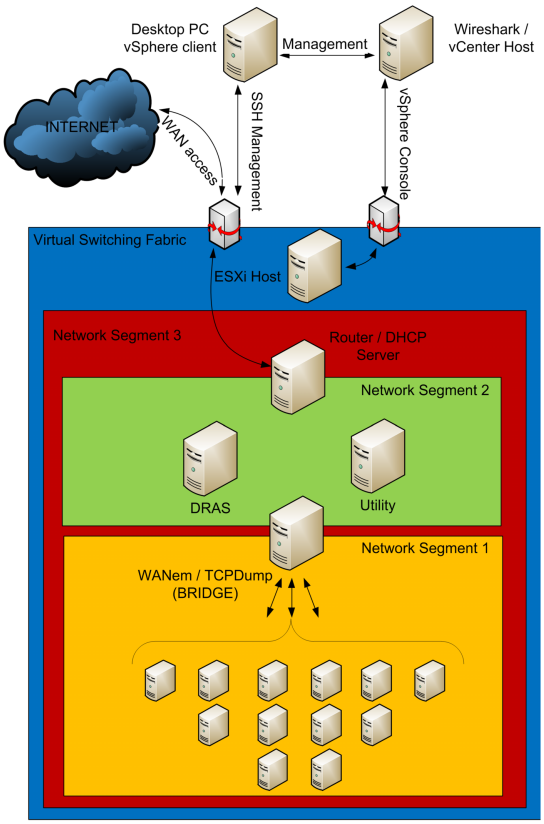


Figure 5. Overview of vNET. S1 represent a NAN. S2 represents a utility's systems. S3 represents the utility's network infrastructure. All segments are software defined and are reconfigurable.

deployed throughout the topology to collect packets at key points.

F. Result capture and analysis TCPDump, Wireshark and ESXi

Utilised computing resources (CPU, Memory, Disk and RAM IO etc.) are recorded using the built in monitoring features of the ESXi host which provides statistics for all virtual machines. This feedback allows development code to be optimised before deployment potentially reducing the end target hardware cost. TCPdump[33] provides a packet level capture of all transmissions and Wireshark [34] provides the facility to analyse this capture in detail for lost and corrupted packets, retransmissions, packet latency and protocol behaviour. This is a key feature for developing distributed application code. It allows code to be certified as functional before deployment.

G. PowerShell scripts

A method of managing this network of virutal machines was needed. This was provided through Windows PowerShell which is a task automation and configuration framework. The VMWare vSphere CLI can be controlled using custom PowerShell. Therefore, to take advantage of the management interface, a number of custom scripts and a CLI was developed to facilitate the following operations:

1) Begin executing the chosen scenario.
2) Set Maximum Transmission Unit (MTU) of specific nodes. This is useful for emulating transactions with increased protocol overhead.
3) Clone VMs from template. Useful for deploying new scenarios rapidly.
4) Destroy VMs.
5) Power up all VMs.
6) Set CPU limit with MHz resolution. Useful for emulating CPU constrained devices.
7) Set CPU reservation. Useful for guaranteeing certain nodes specific processing resources.
8) Power on specific VMs.
9) Restart specific VMs.
10) Change VM network host name.
11) Issue custom command. Useful for providing additional configuration flexibility - takes standard BASH commands and passes them to the chosen nodes.

These scripts allow easy manipulation of the emulation test bed and the virtual machines within it. These files are currently being made publicly available and in the mean time can be requested by email.

III. vNET Validation: Experimental Scenarios

For these experimental scenarios the Message Queue Telemetry Transport (MQTT) [35] protocol was chosen. MQTT is designed to be an efficient, broker-based, publish / subscribe transport protocol. MQTT was chosen to demonstrate vNET as it is well suited to environments with constrained resources, for example where the network is expensive or the embedded devices involved are CPU or RAM constrained. This, coupled with the existing IBM performance [36] analysis will provide a good base for vNET validation.

Few performance analyses have been attempted for MQTT to date [36], [37] and neither has attempted a fully emulated testing approach. Fenton [36] used physical server grade hardware to analyse the performance of the IBM MQTT broker software under load but used artificial traffic generation techniques to emulate incoming traffic which limited the scope of the performance analysis to the broker. Perez took a fully simulated approach using the OMNeT++ [38] network simulator, which restricts the flexibility and limits usefulness of the test-bed. The experiments described here take the fully emulated approach with the addition of real hardware at critical points in the network topology in order to evaluate HIL potential. Unlike the experiments performed by Fenton which focused solely on the broker, the whole network is emulated all the way from broker to clients.

The IBM performance analysis of MQTT [36] specified two scenarios. These have been replicated (but scaled down) in order to show the emulator it can be used to experiment with arbitrary scenarios and network topologies. They are:

1) Multi-publisher, single-subscriber.
2) Multi-publisher, multi-subscriber.

In the first scenario messages are sent to the broker at a rate of 100 per second from randomly selected clients. A single subscriber receives all of the publishes.

In the second scenario each client subscribes to one single topic. Message rates are the same as in scenario 1. It also publishes to a different single topic. Each topic a client subscribes to is published to by only one other client. Therefore,

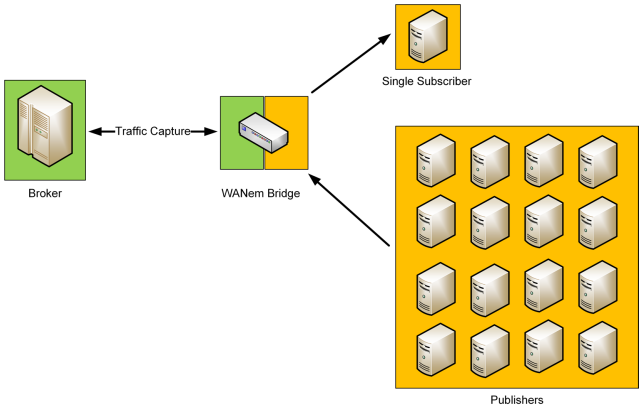


Figure 6. Scenario 1. The broker bridge link uses HIL and brings the traffic into the real world using physical network adapters.

this scenario can be seen as each client having a one to one mapping with another client.

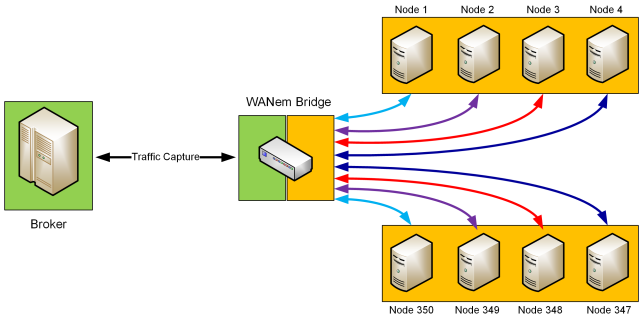


Figure 7. Scenario 2. HIL is used in the same position.

For each scenario all three MQTT QoS levels are experimented with.

1) QoS 0 - The broker/client will deliver the message at most once, with no confirmation.
2) QoS 1 - The broker/client will deliver the message at least once, with confirmation.
3) QoS 2 - The broker/client will deliver the message exactly once, with confirmation.

These six experiments provide results on CPU and network utilisation for both broker and clients.

While the IBM experiments have influenced the scenarios tested in this work, we are are not simply replicating them and there are several key differences. The experiments are scaled to 350 nodes. This is the limit of the hardware being used. Additonal VM hosts would be needed to increase the number. Open source mosquitto [39] is used instead of the IBM Websphere suite. Mosquitto has a smaller resource overhead than the IBM Websphere making it more suited to embedded environments. The clients generating the traffic are individually fully emulated; there is no use of Telemetry Device Daemons to increase the traffic as with the IBM experiments. Since the aim of this experiment is not to simply benchmark the broker, emulating all the clients provides new results which will be used to further analyse MQTT QoS levels and validate the test-bed. The payload for the MQTT messages is a 717 byte XML file that represents a metering update message. The syntax has been borrowed from the Zigbee SEP [40]. Each node client is connected to the broker sequentially. This will allow the CPU time and network bandwidth consumption to be monitored as the number of nodes is increased from 1 to the maximum node number for the experiment. Once the number of nodes reaches this maximum the simulation runs for 15 minutes and automatically shuts down. The network connection between the broker and the software Ethernet bridge utilises a real hardware Intel gigabit ET dual port server adapter looped back on its self. This presents a proof-of-concept for HIL in the test-bed.

Further experimental scenarios based on OpenADR [31] have been performed using the test-bed and the results and analysis of these have been published in *DIRECTOR: A Distributed Communication Transport Manager for the Smart Grid* [5].

A. *vNET Validation: Results*

Figures 8 and 9 show a predictable increase in both CPU usage and bandwidth consumption as the QoS level is increased from 0 to 2.

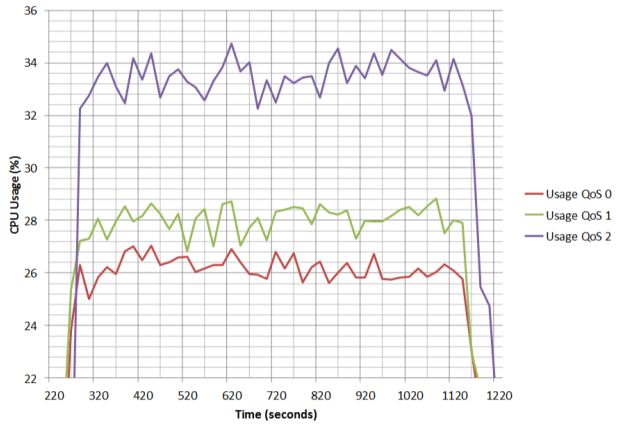


Figure 8. Scenario 1 Broker CPU usage

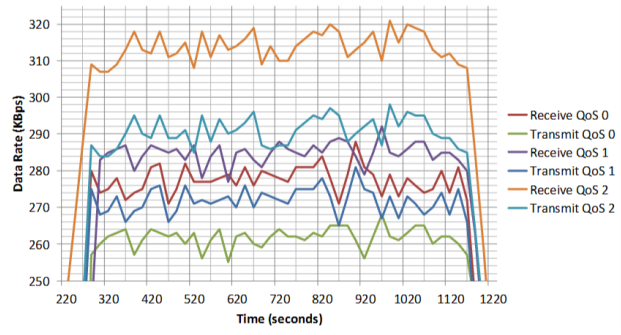


Figure 9. Scenario 1 Broker Network Usage

It can be noted that QoS level 2 is significantly more expensive in terms of processing and networking resources than the previous two levels. This can be seen in Figures 8, 9, 10, and 11. For comparison the average CPU usage for a publishing node in this scenario has been given in Table II.

Figure 10 shows that at most the single subscriber is using approximately 6 times the CPU resources the publishers are

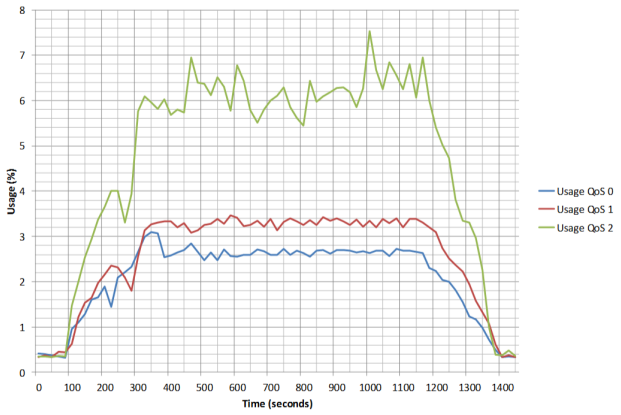


Figure 10. Scenario 1 Single subscriber CPU usage.

even though the network utilization is as much as 360 times as large (The average network utilization for the publishing nodes was between 0.5 and 1 KBps).

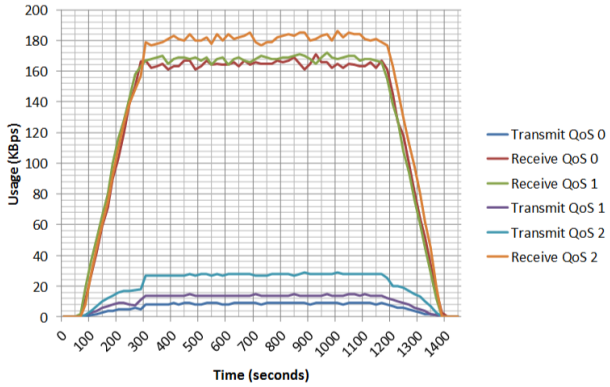


Figure 11. Scenario 1 Single Subscriber Network Usage

Overall the results show that while increasing the QoS level has a significant effect of the resources consumed by the broker, the effect is much less apparent on the individual subscribing nodes. The results from the single subscriber experiments show that as the number of connections increases the cost of higher QoS levels also increases. All results show that the jump from QoS 1 to 2 is much larger than QoS 0 to 1 in terms of processing and networking resources.

TABLE II. CPU usage of the nodes and the host

MQTT QoS Level	CPU per Node (%)	CPU VM Host (%)
0	0.971	69.0
1	0.977	69.7
2	1.022	71.1

B. Scenario 2 - Multi-publisher, multi-subscriber

Figure 12 shows that changing the network topology had little effect on the broker's CPU usage compared to Figure 8. Figure 13 however, does show a significant difference.

In this scenario the data received rate matches the data transmitted rate almost exactly where as in scenario 1 Figure

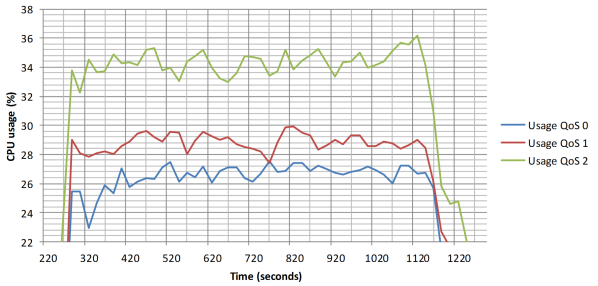


Figure 12. Scenario 2 Broker CPU Usage

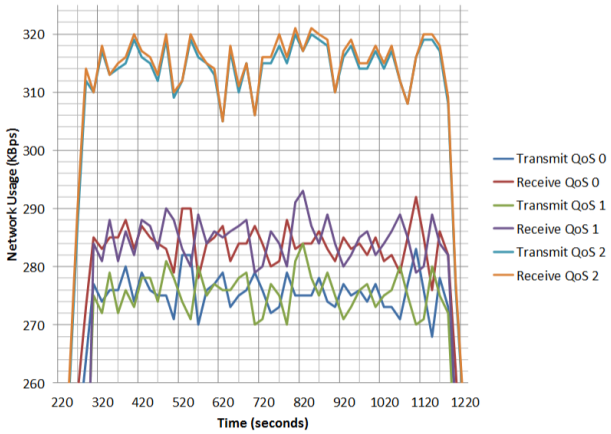


Figure 13. Scenario 2 Broker Network Usage

9 shows that the receive rate was always significantly higher than the transmit rate for a given QoS level. This is due to the single subscriber being overwhelmed by the large flow of traffic being directed at it. Where as in the second scenario this load is distributed across a much larger number of virtual CPUs.

C. Packet Level Analysis

Using TCPdump to collect the traffic passing over the bridge and then Wireshark to analyse it, the following results were obtained.

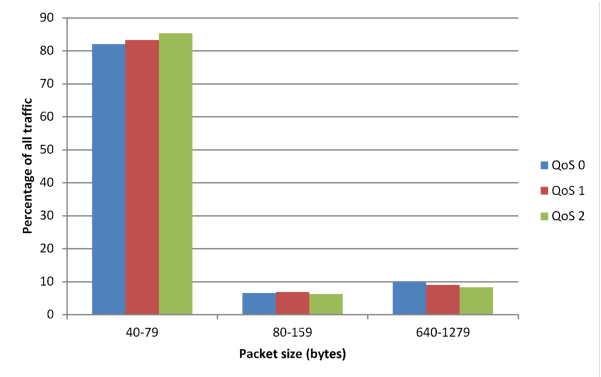


Figure 14. Scenario 1 Distribution of packet sizes

Figure 14 shows that as the QoS level is increased the distribution of packets sizes shifts to having a larger number

of smaller packets and fewer larger ones. The packets lying in the 40-79 byte range are either ARP packets (less than 0.05%) or TCP control messages such as individual ACK, SYN, RST, FIN packets. The 80-159 bytes range are MQTT control packets that have very small payloads such as CONNECT, DISCONNECT or SUBSCRIBE etc. The 640-1279 range are exclusively the PUBLISH packets containing the example XML data. This indicates the large majority of packets are actually TCP control packets, i.e., over 80% of all packets regardless of QoS level are TCP overhead.

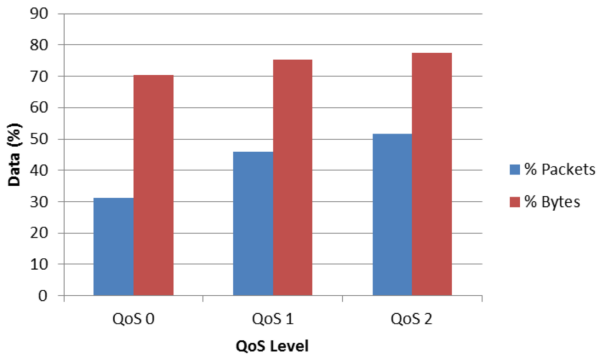


Figure 15. Scenario 1 Proportion traffic that is TCP payload (goodput)

Figure 15 also supports this analysis. At QoS level 1 only 30% of packets actually had a data payload. This includes all of the MQTT packets as well as a large number of PSH-ACK packets with 2-4 byte payloads.

Finally, the use of the Intel Gigabit ET Dual Port Server Adapter presented no noticeable issues. The driver included in the Linux 2.6.32 kernel showed no compatibility issues and its use validated the HIL facility of the test bed.

The results presented are as would be expected. MQTT Level 2 consumes the most bandwidth and CPU time, followed by Level 1 and then Level 0 consumes the least. The results do however provide a valuable step in validating the emulation test-bed. The test-bed was able to produce accurate and detailed results which can actually be used to comment on the appropriateness of using MQTT in resource constrained environments. The main issue is that MQTT operates on top of TCP and that the MQTT replicates large portions of TCP functionality. This means that for DRE environments, MQTT is arguably not well suited. For example, QoS 0 is supposed to be for non-critical messages. There is no confirmation or guarantee that this message is received at the application level. But due to TCP, at the transport level, there inherently is. Normally a low priority, low QoS level message, should provide a low overhead method of communicating. However, due to operating on top of TCP there is in fact a very good chance that the message will be delivered due to TCP's inherent reliable transmission mechanisms. This reduces the usefulness of all the application level QoS levels. Low priority messages do not benefit from as much overhead reduction as they could and high priority messages are partially redundant and overhead is increased with little gain. If MQTT were to not operate over TCP exclusively some transactions would be more efficient in terms of network resources and therefore more appropriate for resource constrained environments. For example, UDP with no reliability features and therefore low overhead, would seem to be a better choice for MQTT QoS 0.

IV. Multi-Protocol QoS: Experimental Scenarios

Using the validated flexible emulation test-bed, vNET, the network topology shown in Figure 16 was configured. The topology is a simple fan out type network where one node is distributing data to a group of 300 nodes representing consumer smart meters.

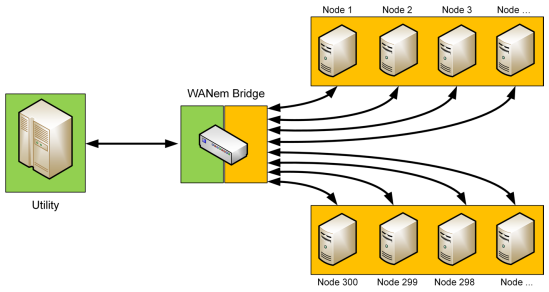


Figure 16. Test bed network topology. 300 nodes are connected to a utility system through a software Ethernet bridge. The Utility publishes the update.

The topology represents the logical grouping that could be used in a Real Time Pricing DSM operation [41]. Traffic shaping is provided by WANem [32], which is a software wide-area network emulator. It provides the ability to manipulate many common network characteristics including available bandwidth, latency and packet loss. The network connection between the utility system and the software Ethernet bridge utilises a real hardware Intel gigabit ET dual port server adapter looped back on its self. This presents a proof-of-concept for HIL in the test-bed.

A. Selected Communication Paradigms

Four viable interaction models were chosen to experiment with; these are shown in Tables III and IV. All are tested with both ideal and resource constrained, lossy network conditions in a set of eight experiments.

TABLE III. - Downlink (utility to consumers) transport choices

Scenario	Transport Protocol	Messaging Pattern
1	TCP	Router / Dealer
2	TCP	Publish / Subscribe
3	PGM	Publish / Subscribe
4	UDP	Request / Response

TABLE IV. Uplink (consumers to utility) transport choices

Scenario	Transport Protocol	Messaging Pattern
1	TCP	Request / Response
2	TCP	Request / Response
3	TCP	Request / Response
4	UDP	Request / Response

Router / Dealer is a tightly-coupled request-response style messaging pattern belonging to the ZeroMQ [42] socket API. It allows messages prefixed with a globally unique identifier (GUID) to be routed to a socket, remote or local, which has that same GUID. Each message sent needs to be prefixed with a valid GUID of a node, which requires additional initialisation

steps in order to acquire this information. Publish / Subscribe is a loosely coupled data distribution style messaging pattern. A publisher publishes a message prefixed with a topic / channel identifier. Only subscribers which have confirmed their interest in messages belonging to this topic / channel get the message routed to them. Scenarios 2 and 3 both use publish / subscribe but they use different transport protocols. Scenario 2 uses standard TCP. TCP is a unicast transport which implies that if a pricing update is to be sent to 300 nodes the Utility node will have to generate and send 300 individually addressed packets (assuming no fragmentation). Conversely, Pragmatic General Multicast [43] (PGM) is an experimental IETF (Internet Engineering Task Force) transport protocol designed to provide reliable multicast communications. In this case, the utility generates only a single packet (again assuming no fragmentation). Whereas TCP and PGM are both reliable transport protocols, UDP is unreliable. It does not have any mechanisms for ensuring reliable delivery but this does mean that it exhibits a lower network overhead.

B. Link configurations of selected network scenarios

Table V shows the network condition scenarios used in conjunction with the scenarios shown in Tables III and IV.

TABLE V. Network Conditions

Network Condition	Description
Ideal	No restrictions on bandwidth (10Gbps nominal effectively unlimited) or any additional latency or packet loss
Resource Constrained	Bandwidth limited to 250Kb/s, additional 30ms +/- 5ms latency and 30% packet loss.

The resource constrained experiment emulates specific network conditions and represents a hypothetical resource-constrained lossy network on the link from the consumers to the utility such as an IEEE 802.15.4 based solution. Even though this represents two opposite extreme scenarios the results would still support the conclusions made for other network conditions. Further experimental details are:

1) In all experiments, the application layer maximum transmission unit (MTU) was configured for each transport protocol to ensure the packet size on the wire did not exceed 127 bytes. This was done to emulate the larger transport overhead (due to fragmentation) that would be seen when using these transport protocols with data link layers that can only support small packet sizes.
2) The virtualised Ethernet bridge interface cards were configured for half duplex communication in order to emulate a half-duplex radio link.
3) The payload used was a 1699 byte Extensible Markup Language (XML) string which is compatible with the OpenADR EventState.xsd XML schema [44].
4) A price update was issued every 0.15 seconds in the request / response architectures and every 45 (0.15*300 = 45) seconds for the publish / subscribe architectures. This approach produces comparable test results as the fundamental differences in how the data is distributed between request / response and publish / subscribe would otherwise make this difficult. All scenarios achieve the goal of generating the same total number of responses from the consumers.
5) All experiments issued price updates for up to 90 seconds and generated 600 responses from the consumers. Tests were allowed to run until all inflight responses were obtained.

The frequency of the Real Time Pricing (RTP) update is higher than any real world application. However, as the number of packets being generated, and hence the congestion, vary linearly with the RTP update frequency, using this frequency simply allows results to be collected easier. The higher frequency has no effect on the conclusions that are made in these experiments.

V. Multi-Protocol QoS: Experimental Results

The results in this section use the UDP scenarios as a baseline. The raw results are shown in Table VI. This is done due to the UDP scenarios representing the simplest combination being experimented with. By using this scenario as benchmark it is easier to see how the other combinations perform in the given network topology against a well understood, ubiquitous transport protocol.

TABLE VI. The raw UDP results that can be used for comparison when results are shown as percentage increases.

	Ideal	Constrained
Utility Data	1.424 MBytes	1.424 MBytes
Consumer Data	1.424 MBytes	1.202 MBytes
Overhead	586.080 KBytes	586.080 KBytes
Message Round Trip Delay	2.183 ms	46.814 ms
Message Loss	0 %	39.3 %

TABLE VII. Message Latency Round Trip Delay (RTD) and Message Loss (PS: Publish / Subscribe, RD: Router / Dealer, RR: Request / Response)

Experiment Number	Message Round Trip Delay (RTD) (ms)	Message Loss (%)
1. TCP PS Ideal	345.6	0.00
2. TCP PS Constrained	21500.0	0.33
3. TCP RD Ideal	5.4	0.00
4. TCP RD Constrained	604.3	0.00
5. PGM PS Ideal	363.7	0.00
6. PGM PS Constrained	17040.0	0.33
7. UDP RR Ideal	2.2	0.00
8. UDP RR Constrained	46.8	39.30

Under the lossy, congested network conditions it took on average 17.04 seconds to complete a round trip for the PGM experiment (Experiment 6) and 21.5 seconds for the TCP (Experiment 2). This is extremely high and is due to the way the consumers are responding; the PGM and TCP publish / subscribe consumers use the same TCP request response architecture to respond with. There is no rate limiting which is causing a large amount of congestion. PGM provides an interface to limit the multicast data rate which would be very useful in this case. It can also be seen that even under perfect network conditions, rate unlimited publish / subscribe architectures are not suitable for applications requiring low latency as individual message delays are over 150 times that of the UDP case (Experiment 1 and 5 vs. 7). The results indicate that the publish / subscribe architectures need a mechanism for rate limiting publishes and responses. The congestion generated when a published pricing update is sent and then responded to

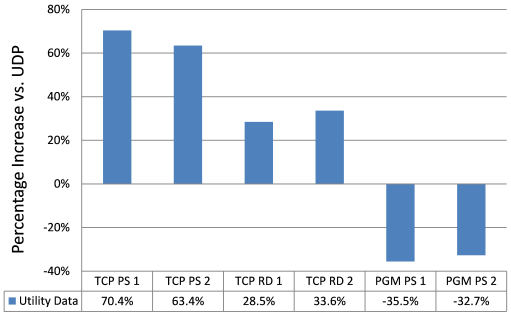


Figure 17. Data sent by the Utility system compared to the UDP scenario.

by all of the consumers simultaneously quickly overwhelms the resource constrained network generating message losses, which in turn cause retransmissions which contribute to the large amount of data generated. This can be seen in Figures 17 and 18. The TCP publish / subscribe scenario shows this better than the PGM scenario. In this scenario the resource constrained, lossy network test actually performs better than the ideal case as the artificially imposed packet delay is having the effect of limiting the packet rate which even with the 30% packet loss and the retransmissions this would introduce, causes the scenario to generate less traffic than the ideal case. This also indicates that a component in the virtual network is being stressed to the point of packet loss under the high packet rates being generated by the low MTU. Even with acknowledging this it shows that high messages rates with relatively large payload compared to the MTU will cause worse congestion problems than 30% packet loss does.

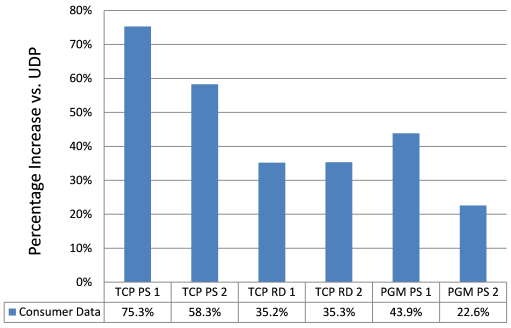


Figure 18. Data sent by the consumer nodes compared to the UDP scenario.

The latency (Tables VI, VII, Experiment 7 and 8) and overhead (Figure 19) results show the UDP experiments outperform both the TCP or PGM equivalents with the TCP. This is not unexpected given the TCP and PGM are both reliable transports, with retransmissions that introduce increased delays against UDP. The notable observation is the performance gap between them. TCP is a generic transport capable of serving many different application requirements quite adequately, but the overhead involved in being so generic is clearly shown in these experiments. There is a clear opportunity to bridge this large gap with a number of UDP based messaging patterns, both unicast and multicast, and apply various application layer reliability mechanisms to them. This would allow applications access to a range of communications service combinations at a higher granularity, so that applications can get a communication service with only the features they need and avoid the general overhead of a one-transport-fits-all approach. Figure

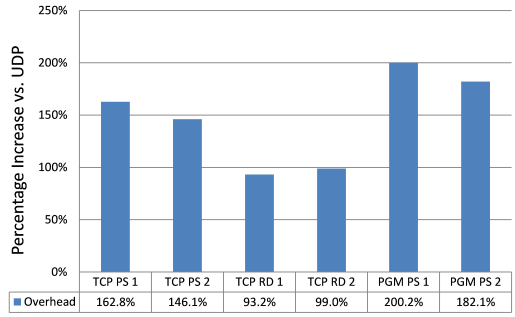


Figure 19. Protocol overhead compared to the UDP scenario measured as any data that was not the XML payload. (Percentage increase is shown)

19 shows a large PGM percentage overhead. Given the much lower overall bandwidth consumed using multicast, this is to be expected. Overhead is calculated as any bytes put onto the wire that are not part of the XML payload. In order to generate the 600 responses (the experimental scenario criteria) from the consumers the utility only has to generate 2 PGM packets (ignoring fragmentation due to the low MTU). The overhead is almost entirely due to the TCP request response uplink from the consumer to the utility. Normalising these results against the total data exchanged shows the PGM architectures are in fact the most efficient next to the UDP architectures. The results show that even though TCP publish / subscribe would appear to be a potential choice for this type of scenario (data distribution with a large fan-out) given that on face value it appears to provide the necessary interface for providing efficient data distribution, it actually performed the worst. TCP-based publish / subscribe involves a high amount of overhead to effectively allow a unicast architecture to emulate services that require a multicast architecture in order to operate efficiently. It provides no network orientated benefit over TCP Router / Dealer. The only benefit it provides is the ability to distribute messages at a more abstract level due to the use of topics / channels. In fact the lack of control on the distribution rate of the messages means that TCP router / dealer is more flexible and consistently generates less overhead and congestion as can be seen in Figures 17-19.

VI. CONCLUSIONS

This paper has presented and validated an argument for exploiting the performance gains achievable by specifically selecting application appropriate transport protocols dynamically at runtime based on specific application requirements. Given the varied network requirements demanded by SG applications and DRE applications in general this approach provides previously inaccessible optimisation opportunities. Furthermore, these gains are achievable without the need to perform costly modifications to any intermediate network infrastructure and would only require modifications to existing networked applications network interfacing code. The cost of this modification could be mitigated by using a middleware system for managing the transport selection. To summarise, the results have shown:

1) For an ideal RTP update distribution use case PGM publish / subscribe and UDP request / response

should be used on the down links and up links respectively for best performance and lowest resource utilisation.

2) For the non-ideal case, the unreliable UDP is only viable if the application can suffer lost responses from the consumer this is a possible scenario. If more reliability is required then another low overhead reliable transport should be used with TCP based options used as a last resort.
3) There is a significant gap between the performance of the Reliable TCP / PGM scenarios and the unreliable UDP scenario in terms of overhead and latency. Additional transports are needed to fill the gap.
4) TCP based Publish / Subscribe provides no network level benefits.
5) Rate unlimited Publish / Subscribe is not viable for applications with a low latency requirement. The packet rate needs to be limited at the point of transmission in order to ensure congestion is not generated.

A large number of additional supported transport protocols, would make it possible for a system to generate custom network interfaces for a much wider range of scenarios in order to improve application performance through manipulation at the transport level. Future work will consider how to automatically manage the large number of potential transport protocol choices which are being suggested using middleware solutions. The virtualisation based network emulation test-bed has been shown to produce useful and viable results. The unique features that the test-bed offers has provided the level of detail normally reserved for scaled real world trials but at a much lower cost. Collecting individual CPU, bandwidth and packet statistics for each node and the overall system has been shown to be quick and convenient. Reconfiguring the test-bed for different topologies can be achieved by consuming much less time and resources compared to a scaled real world trial. Additionally a hardware-in-loop (HIL) proof-of-concept has been presented. The ease at which the technology allows hardware to be assigned and used natively by any virtualized node presents a powerful hardware / driver evaluation tool. The other tools highlighted in this work cannot provide a feature set optimised for the IoT and other distributed, real-time and embedded environments. Therefore, the development of vNET has been justified.

REFERENCES

[1] J. Wilcox, D. Kaleshi, and M. Sooriyabandara, "Multi-Protocol Transport Layer QoS: A Performance Analysis for The Smart Grid," in *IARIA ENERGY 2014, The Fourth International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies*, pp. 13–18.

[2] C. Fraleigh, S. Moon, B. Lyles, C. Cotton, M. Khan, D. Moll, R. Rockell, T. Seely, and S. C. Diot, "Packet-level traffic measurements from the sprint ip backbone," *Network, IEEE*, vol. 17, no. 6, pp. 6–16, 2003.

[3] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications." RFC 3550 (INTERNET STANDARD), July 2003. Updated by RFCs 5506, 5761, 6051, 6222, 7022.

[4] Internet Assigned Numbers Authority, "Protocol Numbers." http://www.iana.org/assignments/protocol-numbers/protocol-numbers.xml. Accessed: 24/03/14.

[5] J. Wilcox, D. Kaleshi, and M. Sooriyabandara, "Director: A distributed communication transport manager for the smart grid," in *Communications (ICC), 2014 IEEE International Conference on*, pp. 4227–4232, June 2014.

[6] A. Mohsenian-Rad, V. W. S. Wong, J. Jatskevich, R. Schober, and A. Leon-Garcia, "Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid," *IEEE Transactions on Smart Grid*, vol. 1, no. 3, pp. 320–331, 2010.

[7] A. Gabaldon, A. Molina, C. Roldan, J. Fuentes, E. Gomez, I. Ramirez-Rosado, P. Lara, J. Dominguez, E. Garcia-Garrido, and E. Tarancon, "Assessment and simulation of demand-side management potential in urban power distribution networks," in *IEEE Bologna Power Tech Conference Proceedings, 2003*, vol. 4, pp. 5 pp. Vol.4–, June 2003.

[8] M. LeMay, R. Nelli, G. Gross, and C. Gunter, "An Integrated Architecture for Demand Response Communications and Control," in *Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, pp. 174–174, Jan 2008.

[9] C.-L. Su and D. Kirschen, "Quantifying the Effect of Demand Response on Electricity Markets," *Power Systems, IEEE Transactions on*, vol. 24, pp. 1199–1207, Aug 2009.

[10] Z. M. Fadlullah, M. M. Fouda, N. Kato, A. Takeuchi, N. Iwasaki, and Y. Nozaki, "Toward intelligent machine-to-machine communications in smart grid," *Communications Magazine, IEEE*, vol. 49, no. 4, pp. 60–65, 2011.

[11] U. Mutlu, R. Edwards, and P. Coulton, "Qos aware bluetooth middleware," in *Information and Communication Technologies, 2006. ICTTA '06. 2nd*, vol. 2, pp. 3239–3244.

[12] M. Henning, "The Rise and Fall of CORBA," *Queue*, vol. 4, no. 5, pp. 28–34, 2006.

[13] Z. Weishan, K. M. Hansen, J. Fernandes, Schu, x, J. tte, and F. M. Lardies, "Qos-aware self-adaptation of communication protocols in a pervasive service middleware," in *Green Computing and Communications (GreenCom), 2010 IEEE/ACM Int'l Conference on Cyber, Physical and Social Computing (CPSCom)*, pp. 17–26.

[14] K. Young-Jin, L. Jaehwan, G. Atkinson, K. Hongseok, and M. Thottan, "SeDAX: A Scalable, Resilient, and Secure Platform for Smart Grid Communications," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 6, pp. 1119–1136, 2012.

[15] Z. Wenjie, Y. Nanhua, C. Hui, H. Jiajian, L. Chuanjian, and L. Xin Jie, "Providing adaptive QoS for Real-Time Publish-Subscribe Service in SGIOC-HQ," in *Innovative Smart Grid Technologies - Asia (ISGT Asia), 2012 IEEE*, pp. 1–5.

[16] B. Zieba, M. v. Sinderen, and M. Wegdam, "Quality-constrained routing in publish/subscribe systems," 2005.

[17] R. E. Schantz, J. P. Loyall, C. Rodrigues, D. C. Schmidt, Y. Krishnamurthy, and I. Pyarali, "Flexible and adaptive QoS control for distributed real-time and embedded middleware," 2003.

[18] Scalable Network Technologies, "Qualnet." http://web.scalable-networks.com/content/qualnet, 2012. Accessed: 24/03/14.

[19] G. F. Riley and T. R. Henderson, "The NS-3 Network Simulator Modeling and Tools for Network Simulation," in *Modeling and Tools for Network Simulation*, ch. 2, pp. 15–34, Springer Berlin Heidelberg, 2010.

[20] S. Doshi, U. Lee, R. Bagrodia, and D. McKeon, "Network design and implementation using emulation-based analysis," in *Military Communications Conference, 2007. MILCOM 2007. IEEE*, pp. 1–8.

[21] K. Fall, "Network emulation in the vint/ns simulator," in *IEEE International Symposium on Computers and Communications, 1999. Proceedings.*, pp. 244–250.

[22] S. Maier, A. Grau, H. Weinschrott, and K. Rothermel, "Scalable network emulation: A comparison of virtual routing and virtual machines," in *12th IEEE Symposium on Computers and Communications, 2007. ISCC 2007.*, pp. 395–402.

[23] T. Song, S. Wen-Zhan, D. Qifen, and T. Lang, "Score: Smart-grid common open research emulator," in *IEEE Third International Conference on Smart Grid Communications (SmartGridComm), 2012*, pp. 282–287.

[24] B. Lantz, B. Heller, and N. McKeown, "A network in a laptop: Rapid prototyping for software-defined networks," in *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*, Hotnets-IX, (New York, NY, USA), pp. 19:1–19:6, ACM, 2010.

[25] F. Osterlind, A. Dunkels, J. Eriksson, N. Finne, and T. Voigt, "Cross-Level Sensor Network Simulation with COOJA," in *Proceedings 2006*

31st IEEE Conference on Local Computer Networks, pp. 641–648, Nov 2006.

[26] A. Dunkels, B. Gronvall, and T. Voigt, "Contiki. A lightweight and flexible operating system for tiny networked sensors," in *29th Annual IEEE International Conference on Local Computer Networks, 2004*, pp. 455–462, Nov 2004.

[27] "VMware vSphere 5.1 Configuration Maximums." http://www.vmware.com/pdf/vsphere5/r51/vsphere-51-configuration-maximums.pdf. Accessed: 24/03/14.

[28] "Introduction to Mininet ." https://github.com/mininet/mininet/wiki/Introduction-to-Mininet. Accessed: 24/03/14.

[29] VMware, "Esxi overview." http://www.vmware.com/products/vsphere-hypervisor. Accessed: 24/03/14.

[30] C. M. Buechler and J. Pingle, *pfSense: The Definitive Guide*. Reed Media Services, 2009.

[31] G. Ghatikar, "Open automated demand response technologies for dynamic pricing and smart grid," *Lawrence Berkeley National Laboratory*, 2010.

[32] H. K. Kalitay and M. K. Nambiar, "Designing WANem : A Wide Area Network emulator tool," in *Third International Conference on Communication Systems and Networks (COMSNETS), 2011*, pp. 1–4.

[33] "Tcpdump." http://www.tcpdump.org/manpages/tcpdump.1.html. Accessed: 24/03/14.

[34] W. Foundation, "Wireshark." http://www.wireshark.org. Accessed: 24/03/14.

[35] D. Locke, "MQ Telemetry Transport (MQTT) V3.1 Protocol Specification." http://public.dhe.ibm.com/software/dw/webservices/ws-mqtt/MQTT_V3.1_Protocol_Specific.pdf, 2010. Accessed: 24/03/14.

[36] O. Fenton, "WebSphere MQ Telemetry V7.0.1 - Performance Evaluation." ftp://public.dhe.ibm.com/software/integration/support/supportpacs/individual/mp0a.pdf, 2010. Accessed: 24/03/14.

[37] J. Perez, "MQTT Performance Analysis with OMNeT++." Networking Insitut Eurecom; IBM Zurich Research Laboratory Switzerland, September 2005.

[38] "OMNeT++, an extensible simulation library." http://www.omnetpp.org/. Accessed: 24/03/14.

[39] R. Light, "Mosquitto. An Open Source MQTT v3.1 Broker." http://linuxcommand.org/man_pages/brctl8.html. Accessed: 24/03/14.

[40] Z. S. Organization, "ZigBee Smart Energy Profile Specification," tech. rep., 01/12/08 2008.

[41] Z. Wei and A. Feliachi, "Residential load control through real-time pricing signals," in *Proceedings of the 35th Southeastern Symposium on System Theory.*, pp. 269–272.

[42] P. Hintjens, "ZeroMQ: The Guide." http://zguide.zeromq.org/, 2010. Accessed: 24/03/14.

[43] T. Speakman, J. Crowcroft, J. Gemmell, D. Farinacci, S. Lin, D. Leshchiner, M. Luby, T. Montgomery, L. Rizzo, A. Tweedly, N. Bhaskar, R. Edmonstone, R. Sumanasekera, and L. Vicisano, "PGM Reliable Transport Protocol Specification." RFC 3208 (Experimental), Dec. 2001.

[44] OpenADR, "OpenADR EventState.xsd XML schema." http://openadr.lbl.gov/src/EventState.xsd. Accessed: 24/03/14.

A Methodology for Accounting the CO_2 Emissions of Electricity Generation in Finland

The contribution of home automation to decarbonisation in the residential sector

Jean-Nicolas Louis, Antonio Caló, Eva Pongrácz
Thule Institute, NorTech Oulu
University of Oulu
Oulu, Finland
e-mails: jean-nicolas.louis@oulu.fi, antonio.calo@oulu.fi, eva.pongracz@oulu.fi

Kauko Leiviskä
Control Engineering Laboratory
University of Oulu
Oulu, Finland
e-mail: kauko.leiviska@oulu.fi

Abstract— **To achieve the decarbonisation of the energy sector in Europe, the CO_2 emission profile of energy consumption must be fully understood. A new methodology for accounting for CO_2 emissions is required for representing the dynamics of emissions. In this article, a dynamic integration of CO_2 emissions due to the electricity production and trade was developed. Electricity consumption and related CO_2 emissions are studied for a typical Finnish household. A model detached house is used to simulate the effect of home automation on CO_2 emissions. Hourly electricity production data are used with an hourly electricity consumption profile generated using fuzzy logic. CO_2 emissions were obtained from recorded data as well as estimated based on monthly, weekly, and daily generated electricity data. The CO_2 emissions due to the use of electric appliances are around 543 $kgCO_2$/y per house when considering only the generated electricity, and 335 $kgCO_2$/y when balancing the emissions with exported and imported electricity. The results of the simulation indicate that home automation can reduce CO_2 emissions by 13%. Part of emission reduction was achieved through peak shifting, by moving energy consumption load from daytime to night time. The paper highlights the role of home automation in reducing CO_2 emissions of the residential sector in the context of smart grid development.**

Keywords-CO_2 emissions calculation, home automation, load shifting, modelling.

I. Introduction

In December 2011, the European Commission set clear goals in its Energy Roadmap 2050 COM(2011)885/2, to achieve a decarbonised society. Decarbonisation in this context means reducing greenhouse gas emissions to 80-95% below 1990 levels by 2050. This will provide considerable challenges for electricity production, consumption and management. Smart grids represent one tool for achieving this target. Smart grids aim at increasing the energy efficiency of the network, peak load shaving, load shifting, and reduction of energy consumption. Smart buildings are expected to be an integral part of smart grids, with smart meters as the gateway allowing the entrance of smartness into the building. Smart meters receive and send information to and from the building for use such as in Home Area Networks, and grid handling. Ultimately, smart buildings will lead to the decarbonisation of the residential sector. A description of CO_2 emissions from the electricity generation in Finland has been presented in our earlier paper [1].

The role of the residential sector in reducing carbon emissions is paramount in the development of the future smart grid [2][3]. A massive deployment of smart meters is under way in Europe, which will facilitate digital measurements, and will allow a consequent access to energy consumption data to energy companies and authorities. Member States of the European Union (EU) have the obligation to implement smart meters covering 80 % of consumers by 2020 at the latest [4]. In contrast to the European Energy Efficiency Directive (2012/27/EU) [4], the Finnish Electricity Market law 588/2013 and its application Act 2009/66 on the electricity supply in the survey and measurement sets the deadline for 2014 [5]. Legal obligations to increase energy efficiency also provide a motivation to the deployment of renewable energy sources (RES) as a vector for energy production, both electrical and heating, in a large scale as well as in buildings. Home energy management systems can have a significant role in contributing to energy efficiency and cutting or shifting peak load. This can be achieved through an active collaboration of energy consuming systems and the information network on a local level [6]. Putting together smart grids, smart buildings, RES-based heat and electricity and energy efficiency, involve the development of a smart energy networks (SEN) capable of managing the energy system through constant monitoring.

The impact of energy efficiency on emissions from the residential sector has been a subject of much research (e.g., [7]-[10]). It has been shown that electric load shifting from the residential sector may reduce air pollution in urban areas [11]. To this effect, developing mathematical tools that are able to anticipate and cut emissions through the deployment of smart systems and home automation is of major importance.

This article aims at exploring the significance of home automation and its impact on the CO_2 emissions of a dwelling, and the possible ways home automation can contribute to decarbonisation. In the first Section of the paper, a description of CO_2 emissions from the production and the use of electricity in Finland will be presented. The second Section presents the methodology used for translating hourly carbon emissions to single households will be described. The third

Section shows and details the results from the simulations carried out on two chosen types of dwellings, which will be described and analysed.

II. Related Research

Research on smart houses and their development has been going on for quite some time. Smart homes can be broadly seen as buildings monitored and controlled for multiple purposes [2]. The energy management feature of smart homes is one aspect of the development. Algorithms for generating electricity consumption load profile have been developed on hourly and half-hourly bases [12], but also with a finer grid on a minute-basis [13]. These algorithms can be further used to emphasise the potential of energy in smart houses and their roles in improving energy efficiency, reducing energy consumption and CO_2 emissions from the energy used. More elaborate algorithms have also been developed, where the integration of each appliance within the dwelling has been modelled with a bottom-up approach [14][15]. Finally, the management of appliances within the dwelling may as well be implemented in simulation for optimizing their usage and enhancing demand-side management [16][17].

Previous studies have attempted to determine the impact of energy efficiency measures on CO_2 emissions from the residential sector [7]-[10]. Detailed algorithms for evaluating CO_2 emissions associated with electronic appliance usage have been proposed [18]. One of the main drawbacks of previous methods is that CO_2 emissions are based on a fixed coefficient, thus limiting the understanding of the CO_2 emission mechanism. The variation of electricity production and market dynamics have been ignored, resulting in a biased estimation of carbon dioxide emissions. A more dynamic model has been elaborated by Stoll et al. for estimating CO_2 emissions and their impact on demand response [19]. Although the research of Stoll et al. has based its dynamism on real dataset of energy production on an hourly basis for various countries, the CO_2 emissions related to the production of electricity are based on a fixed emission factor from the IEA annual report on CO_2 emissions [20]. Therefore, the dynamic has increased but variation due to the use of different fuel types were not present and, therefore, the estimation is severely biased. Fuel usage varies according to market prices, resource availability and climatic variations.

Consequently, studies on segmented electricity production, related CO_2 emissions, and the impact of home automation on the emissions are lacking.

III. Electricity consumption and Carbon Emissions in the residential sector

In terms of CO_2 emission reduction in the residential sector, the largest effort should be made in retrofitting buildings. The average renewal time of the residential sector is estimated to be around 70 years [7][12]. The influence of technology on CO_2 emissions needs to be highlighted. Consequently, technology upgrading can greatly influence the total CO_2 emissions of the residential sector. Lighting consumes over 30 % of the total electricity used in households [13][21]. The upgrade of lighting technology is one way for impacting energy consumption [14][15], but also for reducing carbon emissions [16][17][22]. Furthermore, home energy management systems will continue to play a role for increasing energy efficiency, reducing energy consumption [7][23] and allowing load shifting.

In Finland, electricity generation and consumption is being constantly surveyed, recorded and reported by Statistics Finland. In 2012, household appliances consumed 8 072 GWh of electricity [18][21]. At the same time, 2 579 781 households were registered in Finland [19][24], resulting in an average consumption per house of 3 129 kWh/y. There can be considerable deviation from this average value, if the households is in an apartment building or a detached house [20][25]. Furthermore, total electricity production in Finland was around 67.7 TWh in 2012, while the total consumption of electricity was around 82.9 TWh, and a total of 8.4 $MtCO_2$ were emitted. Therefore, it can be estimated that the share of electricity using devices in the total CO_2 emissions from electricity production and consumption are 1001 tCO_2/GWh_{pro} or 817 tCO_2/GWh_{cons}.

IV. Methodology

Data acquisition consisted of analysing the electricity generation of all power plants in Finland, categorized according to their primary fuel, and the categories of power plants on an hourly basis. Secondly, the CO_2 emissions associated with the aforementioned categories were calculated on an hourly basis. Monthly CO_2 emissions are available from July 2011 to April 2014 [26]. It is then possible to evaluate the CO_2 emissions on an hourly basis by correlating the primary energy source for electricity generation and associated monthly CO_2 emissions.

A. Energy Data Collection

Data on electricity production and production forecast are to be reported to the grid aggregator of the Nordic network. Fingrid, the transmission system operator (TSO) of Finland releases information about the network operation and sources of electricity production on their network. In parallel, TSOs of all Nordic countries must report planned and unplanned interruption to the grid through the Urgent Market Message (UMM) system that allows for a better management of the electric grid. Data information on the electricity production systems have been recorded every 5 minutes from Fingrid. As Fingrid does not provide historical data on electricity production, the missing data are completed by two methods. Between 2010 to 2014, the use of historical daily information on the technology used for producing electricity, combined with the UMM system recorded by NordPool for integrating system failure into the data vector is used. The second method

uses the monthly and weekly information for disaggregating the energy production data at the country level on an hourly basis.

1) Daily Energy

Information on daily power availability is reported by the Finnish Energy Industry Association. The information is split into five categories: nuclear power, combined heat and power (CHP), wind power, hydropower, and separate thermal power. As the characteristics of these technologies are different, the following assumptions have been made: nuclear power has a somewhat steady production of electricity; thermal power, which includes CHP electricity production and separate thermal power, have a production of electricity proportional to the total electricity production; and, hydro is used for balancing electricity production. The notations *h*, *w*, *d*, and *m* designate the hourly, weekly, daily, and monthly time step respectively, *i* is the energy technology used for producing the electricity, and *tot* stands for the total amount of a unit countrywide.

$$P_{h,tot} = P_{h,th-CHP} + P_{h,th-ind} + P_{h,nu} + P_{h,wi} + P_{h,hy} \quad (1)$$

Where P_h is the power produced on an hourly basis [MW], and *tot* stands for the total electricity produced, *nu* is nuclear power plants, *th* is thermal power plants, *wi* is wind power, and *hy* is hydropower.

Thermal power consists of CHP units from district heating and industrial sites, and the separate thermal units. Each unit runs proportionally to the total electricity produced balanced by the share of electricity brought by a unit as a ratio of power available. Therefore, the hourly production of electricity from thermal power plants $P_{h,th-i}$ can be written as:

$$P_{h,th-i} = \frac{P_{d,i}}{\sum_{i=1}^{5} P_{d,i}} \cdot P_{h,tot} \cdot \rho \quad (2)$$

Where $P_{d,i}$ is the daily power used as reported by the Finnish Energy Industry Association [27] for the a particular technology [MW], $P_{h,tot}$ is the total electricity production data provided by the TSO [MW], and ρ is the ratio of the weekly energy produced between the district heating and the industrial electricity produced [‘]. In case of separate thermal power or gas fired turbine, ρ is equal to 1.

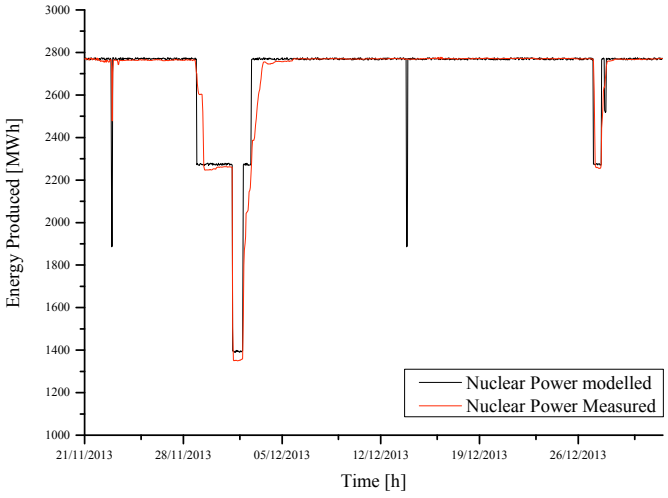


Figure 1. Measured and modelled data of electricity production from nuclear power plants

The flexibility of power production form nuclear power plants is rather low. Consequently, the output power is assumed to be constant with low random fluctuations as expressed in (3).

$$P_{h,nu} = P_{in,nu} \times \left(1 - R \sim U\left(\left[-2.8 \cdot 10^{-4}, 2.8 \cdot 10^{-4}\right]\right)\right) - P_{f,nu} \quad (3)$$

Where $P_{h,nu}$ is the hourly electricity produced by the nuclear park [MWh/h], P_{in-nu} is the global energy produced by the total power of the nuclear park [MWh/h], and P_{f-nu} is the power fault that occurs for each nuclear power station expressed in terms of energy evaluated from the UMM [MWh/h]. The standard variation of steady-state power output on an hourly basis from the nuclear power plants has been calculated from the measured data and is equal to 0.028 %. Fig. 1 illustrates the correlation between the previous equation and the measured data form Fingrid. It can be noticed that the inertia when a power plant is being disconnected has not been integrated, and thus bring a small bias. Notwithstanding, for the purpose of calculating the CO_2 emissions, this approximation is considered acceptable.

Wind power production is based on a generalised model [28] that takes into account the nominal wind power of a station and the characteristics of the wind turbine, and the wind park statues development in Finland as summarised in Table I. Consequently, wind power is calculated for a wind park A, where A is a 1-by-n matrix that varies depending on

TABLE I. WIND ENERGY PRODUCTION SYSTEMS IN FINLAND, BASED ON [29]

Year	WT_{in} [‘]	WP [kW]	WP_{min} [kW]	WP_{max} [kW]	WP_{in} [MW]
2000	63	602	65	1300	37.92
2001	63	614	65	1300	38.7
2002	64	666	200	2000	42.635
2003	74	664.1	200	2000	43.835
2004	92	869	75	2300	79.08
2005	94	898	75	3000	86.215
2006	96	898	75	3000	86.215
2007	107	1028	75	3000	110.015
2008	118	1212	75	3000	143.015
2009	118	1235	75	3000	147.015
2010	130	1475	75	3600	199.115
2011	131	1519	75	3600	198.99
2012	151	1700	75	3600	258
2013	209	1850	75	3600	447
2014	209	1850	75	3600	447

the daily available power for wind turbines WP_d, where n is defined using (4).

$$n = \left\lfloor WT_{in} \cdot \frac{WP}{WP_{in}} \right\rfloor \quad (4)$$

Where *WP* is the daily wind power available for producing energy. Each value of *n* is calculated as a uniformly distributed random number *X*:

$$X \sim U\left(\overline{WP}, b\right) \epsilon \left[WP_{min}, WP_{max}\right] \quad (5)$$

As hydropower is the most flexible type of energy production system, it is assumed that it is capable of producing the remaining energy needed for fulfilling the total electricity production reported by the TSO.

$$P_{h,hy} = P_{h,tot} - \sum P_{h,i} \quad (6)$$

2) Monthly and Weekly energy Data

Monthly and weekly energy data are used in the case where energy production before 2010 needs to be modelled. Before 2010, daily information on energy production systems is not available.

The electricity generated in Finland on an hourly basis is reported by the Finnish Transmission Service Operator – Fingrid since 2004 [30]. The data is split into two groups: electricity generated by power plants and the electrical load on the network taking into consideration the import and export of electricity. Moreover, the Finnish Industry Association (Energiateollisuus) recorded weekly electricity generated from 1990 [31], which is broken down by the technology used: wind, hydropower, nuclear, CHP industry, CHP district heating, conventional and gas turbine power plant. Finally, Fingrid informs in real-time the state of the network, using the same categories as mentioned above. Thus, for building up the hourly electricity generation by categories for the years 2010+, the weekly average electricity production by category is used, in parallel with the hourly electricity generated countrywide. The exported electricity is considered in the electricity generated and in corresponding CO_2 emissions. The imported electricity is considered as a share of CO_2 emissions from electricity consumption in Finland. In order to include the imported electricity into overall emissions from electricity consumption in Finland, it is necessary to know the energy mix for producing the electricity of the country from which Finland is importing. The hourly electricity generated from a particular energy source in the primary country is evaluated using (7).

$$P_{h,i} = \left(\frac{P_{w,i}}{P_{w,tot}} \cdot \frac{P_{in,i} - P_f}{P_{in,i}}\right) \cdot P_{h,tot} \quad (7)$$

Where $P_{h,i}$ is the electric energy generated by a given technology per hour [MWh/h], $P_{w,i}$ is the electric energy generated on a weekly basis by a given technology [MWh/w], $P_{w,tot}$ is the total amount of electricity produced in Finland per week [MWh/w], $P_{h,tot}$ is the total amount of electricity produced per hour [MWh/h], $P_{in,i}$ is the total installed power for the technology *i*, and P_f is the power fault that occur for each power station expressed in terms of energy evaluated from the UMM [MWh/h].

Nuclear power, wind power, and hydropower is evaluated using the same method as the one presented in the daily energy section, except that the weekly energy produced is used instead of the daily power available, as the value of WP in (4).

Once the hourly electricity generated by technology has been defined, it is possible to evaluate the hourly emissions from all the power plants.

B. Emission Data disaggregation

Emissions of CO_2 for electricity production consider only those directly related to the production of electricity: the net and gross emissions. Gross emissions consider only the emissions related to the electricity production within the country, while the net emissions evaluates the balance of emissions due to the import and export of electricity. Therefore, emissions related to fuel transportation or waste management are neglected. The power plants emitting CO_2 and equivalent greenhouse gases are thermal power plants. Thermal power plants can be divided into three distinctive categories in Finland. The first category integrates all power plants primarily used for producing heat for district heating. Electricity is therefore a by-product and varies depending on the thermal power need. The second category includes industries that produce electricity as a by-product from their activity. They may have seasonal variations depending on the industrial activities. The third category includes separate electricity production and groups all the thermal power plants that uses gas turbine, or is used for producing only electricity from thermal power plants. Some of the separate power plants are used for peak load hours, others for aiming at the stability of the grid, or simply to produce electricity. The following sections detail the composition of the conventional power plants in Finland and they are classified following the main fuel type. This description will help at disaggregating data from the energy source used for producing electricity from the above-mentioned three categories.

Data on energy source usage are available on a monthly basis since July 2011. Therefore, two cases are made distinct, the period from July 2011 to 2014 will be processed using the first methodology, and data from 2004 to July 2011 will be processed using the second methodology. The first methodology consists of calculating the monthly energy mix for each technology. This is to integrate the variation of raw material usage in the energy industry. The second methodology considers the measured energy data, the related calculated emissions, and the variation of outside temperature. From these two main variables, it is possible to correlate the variation of energy production and outside temperature to the emissions using a multi-linear regression model.

1) Emissions 2011 - 2014

The first two main categories of technologies that are used to produce electricity as a by-product are the electricity from the district heating, and industrial CHP units. The third category produce electricity during peak load hours or on a permanent basis: separate power plants. Each segment uses different sources of energy that are summarised in the Table II. Also, each segment can be represented in terms of number of units or power capacity. This is to differentiate and understand the emissions levels from the electricity production.

By using the distribution given in Table II and the monthly reported amount of raw energy used, the monthly emissions for each of conventional thermal power plant, separate thermal power plant excluding gas engines, gas power plant, CHP from district heating and CHP from industrial electricity production are calculated with:

$$Em_{m,x} = \left[1 - \frac{\begin{bmatrix} P_x \\ \vdots \\ \vdots \end{bmatrix}}{\begin{bmatrix} P_x \\ \vdots \\ \vdots \end{bmatrix} + \begin{bmatrix} P_y \\ \vdots \\ \vdots \end{bmatrix}}\right] \times \begin{bmatrix} P_{Em_{x+y}} \\ \vdots \\ \vdots \end{bmatrix} \quad (8)$$

Where $E_{m_{m,x}}$ is a n-by-1 matrix of the monthly emissions of the electricity production from district heating or from industry [ktCO$_2$/m], P is the installed power for each raw material where x and y stand for the district heating and the industrial sector [MW] respectively. Fig. 3 is the resulting monthly CO$_2$ emissions from iterating within above equation.

As part of its legal obligation, Finland reports CO$_2$ emissions from power plants, and energy intensive industry [33]. The Finnish Industry Association estimates monthly specific emissions related to electricity production, based on the type of fuel used by the energy industry [31]. By knowing the hourly electricity production from each sector, we can estimate the CO$_2$ emissions for each hour countrywide using (9) to (12).

$$E_{w,i} = a \cdot \left(\frac{P_{w,i}}{P_{m,i}} \cdot \frac{\delta_m}{7} \right) \quad (9)$$

Where a is evaluated using (10) if the full week is within the same month n, or (11) if the full week is between two months, n and $n+1$.

$$a = \frac{7E_{m,n}}{\delta_m} \quad (10)$$

$$a = \left(\delta_w \cdot \frac{E_{m,n}}{\delta_m} \right) + \left(\frac{E_{m,n+1}}{\delta'_m} \cdot (7 - \delta_w) \right) \quad (11)$$

Finally, the hourly emissions are given by,

$$E_{h,i-gen} = P_{h,i} \cdot \frac{E_{w,i}}{P_{w,i}} \quad (12)$$

Where $E_{h,i-i}$ is the emissions from the electricity generated hourly by the given technology [ktCO$_2$/h], and $E_{w,i}$ is the weekly emissions by technology segment [ktCO$_2$/w], δ_w is the day number within a week where Monday is 1 and Sunday is 7, δ_m and δ'_m are the number of days in the studied months, $E_{m,n}$ is the monthly CO$_2$ emissions for the month n. Fig. 2 illustrates the energy generated and its corresponding CO$_2$ emissions on an hourly basis for the year 2012 in Finland. It can be noticed that, although there is a strong correlation of CO$_2$ emissions to electricity generation, emissions may decrease even though the energy generation increases, due to the fact the energy mix is changing Fig. 2.

The emissions due to the electricity imported are added to the primary emissions from the electricity generated within the country. The CO$_2$ emissions from the electricity generated dedicated to the export is then subtracted from the hourly emissions $E_{h,cI}$. In order to account the net CO$_2$ emissions from the electricity load in the country, the emissions from each country with which Finland is trading electricity are evaluated, meaning Norway, Sweden, Russia and Estonia. As the hourly energy mix is not known for each country, a general coefficient of CO$_2$ emissions has been considered for 14 kgCO$_2$/MWh$_{pro}$ for Norway, 21 kgCO$_2$/MWh$_{pro}$ for Sweden, 417 kgCO$_2$/MWh$_{pro}$ for Russia and 1 059 kgCO$_2$/MWh$_{pro}$ for Estonia [20].

The share of CO$_2$ emissions coming from each trading country is evaluated using (13).

$$E_{h,c_n} = \sum_{i=2}^{n} \frac{P_{h,net-c_i}}{P_{h,load}} \cdot E_{c_i} \quad (13)$$

Where E_{h,c_i} is the hourly emissions for each participating country to the electricity trade [kgCO$_2$/h], $P_{h,load}$ is the hourly electric load on the Finnish network [MWh/h], $P_{h,net-c_i}$ is

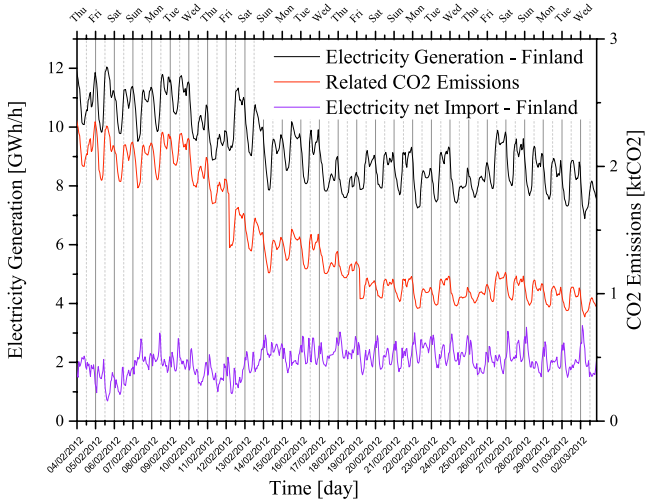


Figure 2. Hourly electricity generation, net import and their related CO2 emissions from 4.02-02.03.2012

TABLE II. INDUSTRIAL PARK PRODUCING ELECTRICITY FROM CHP UNITS IN FINLAND FROM INDUSTRIAL AND DISTRICT HEATING POWER PLANTS, BASED ON [32]

Industry CHP				District Heating CHP				Separate Power Plants			
Ass. Cat	Declared Main Fuel	Nbr of PP	Total Power	Ass. Cat	Declared Main Fuel	Nbr of PP	Total Power	Ass. Cat	Declared Main Fuel	Nbr of PP	Total Power
Peat	Peat	16	643.2	Natural Gas	Natural gas	17	1239.4	Oil	Medium heavy distillates	20	1098.9
Biomass	Industrial wood residues	19	462.6	Peat	Peat	19	1102.24	Peat	Peat	1	154
Natural Gas	Natural gas	10	427.3	Biomass	Forest fuelwood	7	57.7	Oil	Heavy distillates	6	342
Others	Other by-products and wastes used as fuel	1	7	Coal	Hard coal and anthracite	7	1647.1	Oil	Light distillates	3	16.9
Biomass	Forest fuelwood	2	63	Oil	Medium heavy distillates	2	2.9	Coal	Hard coal and anthracite	5	1751
Biomass	Black liquor and concentrated liquors	16	1152.3	Biomass	Industrial wood residues	3	142.5	Coal	Blast furnace gas	2	95.8
Others	Other non-specified energy sources	1	3.9	Others	Other by-products and wastes used as fuel	1	9	Natural gas	Natural gas	2	2.7
Coal	Hard coal and anthracite	1	4.2								
Biomass	By-products from wood processing industry	1	64								
Oil	Heavy distillates	3	13.8	Oil	Heavy distillates	2	177				
Others	Exothermic heat from industry	2	39.3	Others	Biogas	1	14.42				
Oil	Light distillates	1	1								
Others	Biogas	3	4.9								

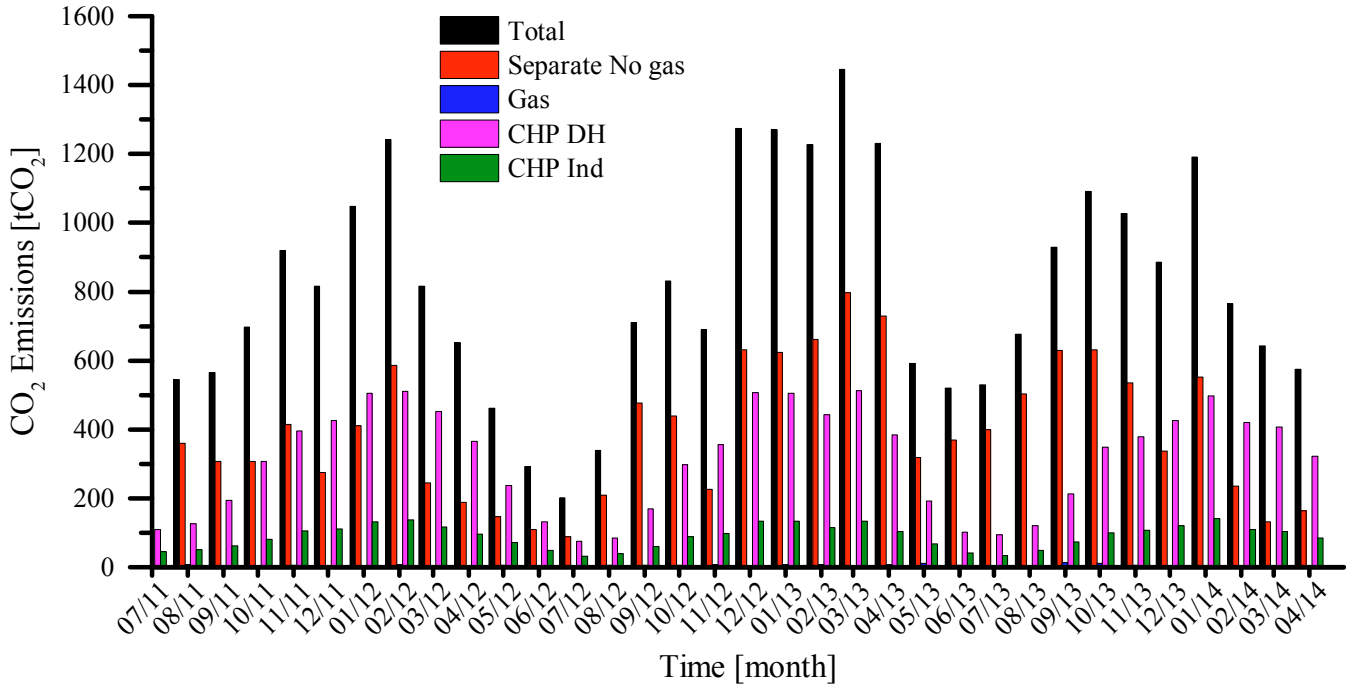


Figure 3. CO_2 emissions in Finland from conventional thermal power plants excluding gas engines, gas power plants, CHP from district heating and CHP from industrial electricity production, based on [32].

the net balance of electricity traded between Finland and the country *n* [MWh/h] in case of export or the difference between electricity generated and the electricity exported in the case of Finland, and E_{c_i} is the coefficient of CO_2 emissions for the corresponding country [kgCO_2/MWh]. In case $P_{h,net\text{-}c_i}$ is negative, the coefficient of CO_2 emissions is equal to $E_{h,i\text{-}gen}$ as the emissions from the Finnish production is exported as well, otherwise, E_{c_i} takes the value defined by the IEA.

Finally, the hourly emissions E_h are determined as the sum of the hourly emissions for each participating country to the electricity trade E_{h,c_i} as shown in (14).

$$E_h = \sum_{i=1}^{n} E_{h,c_i} \tag{14}$$

The emission data in Fig. 2 are then translated to a single house where the hourly electricity consumption profile has been previously generated using (8).

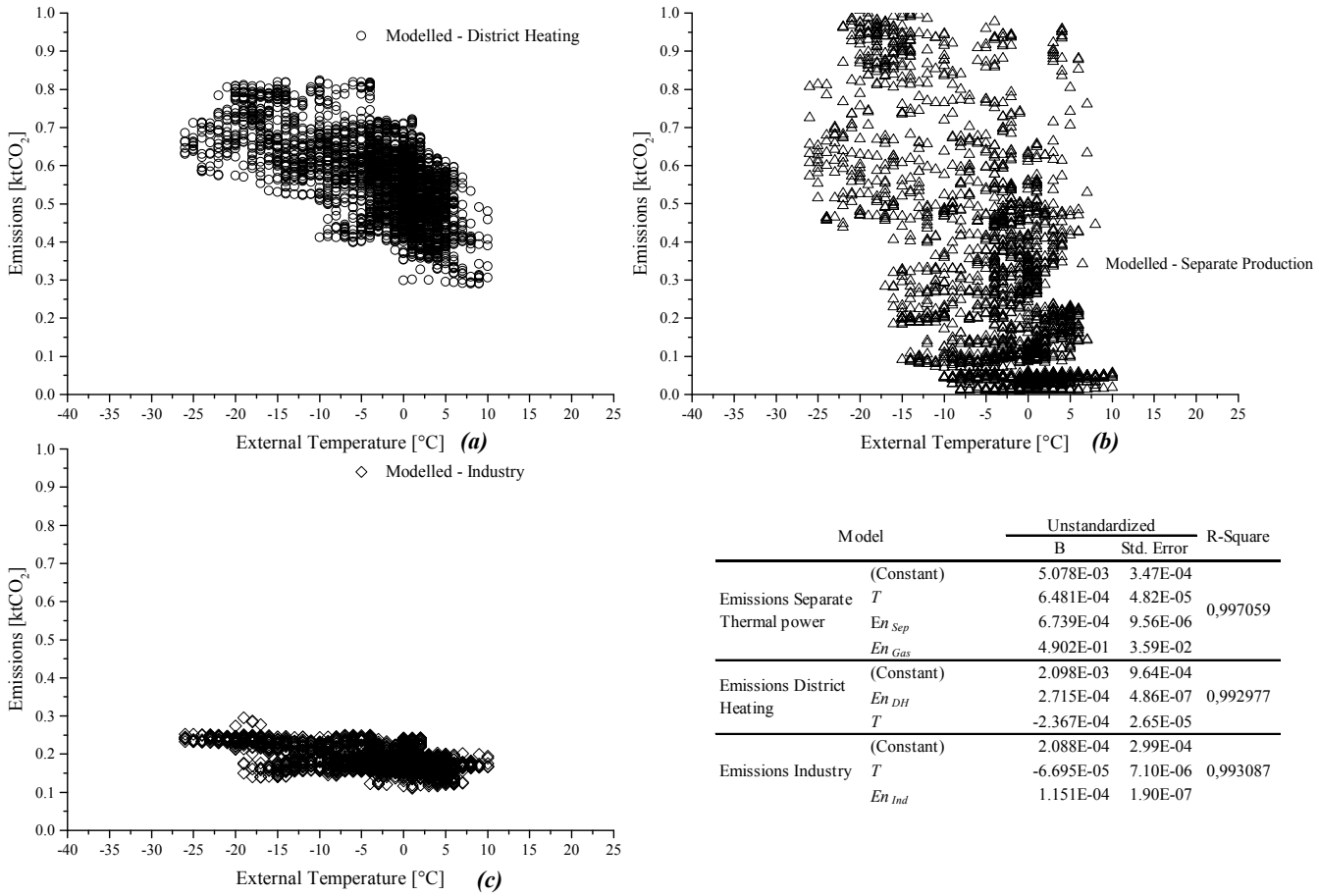
$$E_{h,house} = \sum_{j} \frac{P_{j,house}}{P_{h,tot}} \cdot E_h \cdot 10^3 \tag{15}$$

Where $E_{h,house}$ is the hourly emissions from the house [kgCO_2/h], and $P_{j,house}$ is the total hourly electricity consumed by the house excluding the electric heating [kWh/h]. Two cases are differentiated: CO_2 levels towards the production of electricity within the primary country, and the net CO_2 emissions level considering the import and export. In the first case, *P* takes the value of the total electricity produced in the primary country $P_{h,tot}$. In the second case, *P* takes the value of the total load on the electric grid of the primary country $P_{h,load}$.

The results give an estimate of CO_2 emissions related to the electricity consumption in a private household on an hourly basis. This model is then applied to an average Finnish dwelling previously modelled, in order to estimate the daily CO_2 emissions.

a) Gross emissions from DH

In order to extend the research on multiple years, measured data of electricity production have been considered alongside with the monthly emission data and the resulting hourly emission data generated using the method previously explained. From the hourly CO_2 emissions calculated, the correlation between energy production and external temperature has been evaluated. It appears that the variation of CO_2 emissions and air temperature has a good correlation with a Pearson coefficient R of -0.627 as Fig. 4 and Fig. 5 illustrates. Therefore, the multi-linear regression with a R-square of 0.993 can be written as follows,



Model		Unstandardized B	Unstandardized Std. Error	R-Square
Emissions Separate Thermal power	(Constant)	5.078E-03	3.47E-04	0,997059
	T	6.481E-04	4.82E-05	
	En_{Sep}	6.739E-04	9.56E-06	
	En_{Gas}	4.902E-01	3.59E-02	
Emissions District Heating	(Constant)	2.098E-03	9.64E-04	0,992977
	En_{DH}	2.715E-04	4.86E-07	
	T	-2.367E-04	2.65E-05	
Emissions Industry	(Constant)	2.088E-04	2.99E-04	0,993087
	T	-6.695E-05	7.10E-06	
	En_{Ind}	1.151E-04	1.90E-07	

Figure 4. Calculated emissions from the measured power produced by separate thermal power plant and the multilinear regression used to model the corresponding emissions as factors of the external temperature and the measured power produced.

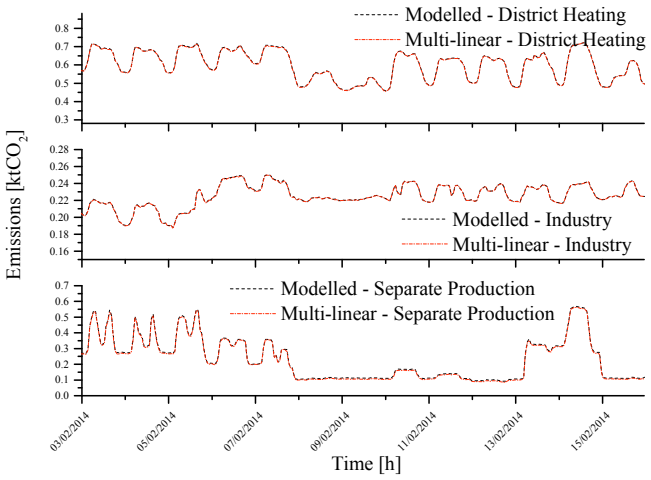


Figure 5. Correlation between the modelled emissions from the energy production level and the multi-linear regression.

$$E_{m_{DH}} = -2.37 \cdot 10^{-4} T + 2.72 \cdot 10^{-4} E_{n_{DH}} + 2.098 \cdot 10^{-3} \quad (16)$$

Where $E_{m_{DH}}$ is the CO_2 emissions from the DH industry [$kgCO_2$], and T is the external temperature [°C].

b) Gross emissions from Industry

Similarly, the emissions from the electricity produced from industrial processes correlate with the variation of external temperature, with a Pearson coefficient of -0.527. The resulting equation considers the variation of external temperature and electricity production level. (17) gives the emission from this industrial segment with a R-square of 0.993.

$$E_{m_{Ind}} = -6.7 \cdot 10^{-5} T + 1.15 \cdot 10^{-4} En_{Ind} + 2.09 \cdot 10^{-4} \quad (17)$$

Where $E_{m_{ind}}$ is the CO_2 emissions from the industrial processes [$kgCO_2$], and T is the external temperature [°C].

c) Gross emissions from separate thermal power

The third segment integrates two types of electricity production technologies: the gas turbine that has a minor role in producing electricity, and condensing power plants using oil, coal, and peat as main fuels that represent the main source of electricity.

$$E_{m_{Sep}} = 6.48 \cdot 10^{-4} T + 6.74 \cdot 10^{-4} En_{Sep} + 0.490195 En_G + 5.078 \cdot 10^{-3} \quad (18)$$

Where $E_{m_{sep}}$ is the CO_2 emissions from the separate thermal power plants [$kgCO_2$], and T is the external temperature [°C].

2) Gross and Net emissions from Finnish electricity

The emissions were detailed by technology at the country level; therefore, it is possible to speak about gross emissions of CO_2 from the electricity production. As it has been mentioned earlier, imported electricity has also been considered in the evaluation of CO_2 emissions from households. Emissions in a country vary daily, weekly and seasonally. Overall, net emissions are always lower than gross emissions due to the fact that Finland imports less CO_2-intensive electricity than it exports. Fig. 6 illustrates the variation of gross and net CO_2 emissions and the deviation between both emissions.

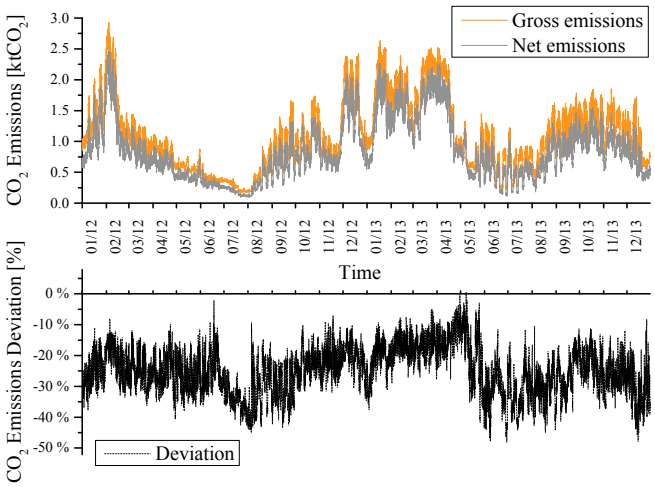


Figure 6. CO_2 emission variations between gross and net emissions including exported and imported electricity.

The deviation between net and gross emissions varies from +0.56 % to -48.13 % with a median value of -24.71 %. This means that the balance of imported and exported electricity is environmentally beneficial for Finland. The import and export mix also varies hour-by-hour. Nevertheless, a trend can be observed on a yearly basis; Finland is importing mainly (97 %) from Sweden and Russia while exports are mainly focused on Estonia and Sweden. Norway, which has the lowest CO_2 emissions factor, plays a minor role in the Finnish electricity mix due to the lack of high voltage transmission line north-south, and the sparse population in Northern Finland.

Depending on how the emissions are accounted, very different results can be obtained from the study of households. Therefore, two cases will be studied; one with gross and another with net emissions.

V. Smart House Emissions

The emissions related to electricity consumption from the residential sector can be determined based on the emissions from the production and trade of electricity, and were estimated hour-by-hour. For this purpose, an electricity consumption model was built for simulating various types of detached houses with multiple configurations such as the number of household members, number and types of appliances with their related energy efficiency factor and so on. A detached house, home to 4-persons has been simulated, with various types of technologies installed in it [34]. Inhabitants are rated depending on their willingness to respond positively to an action. Green users are considered to have a positive response up to 70% of the time, orange users 50% and brown users 30%. This research is focusing on green users while previous results of users' influence on home automation consider all three [34].

A. Electrical and electronic devices in the house

The house electricity demand profile is drawn on an

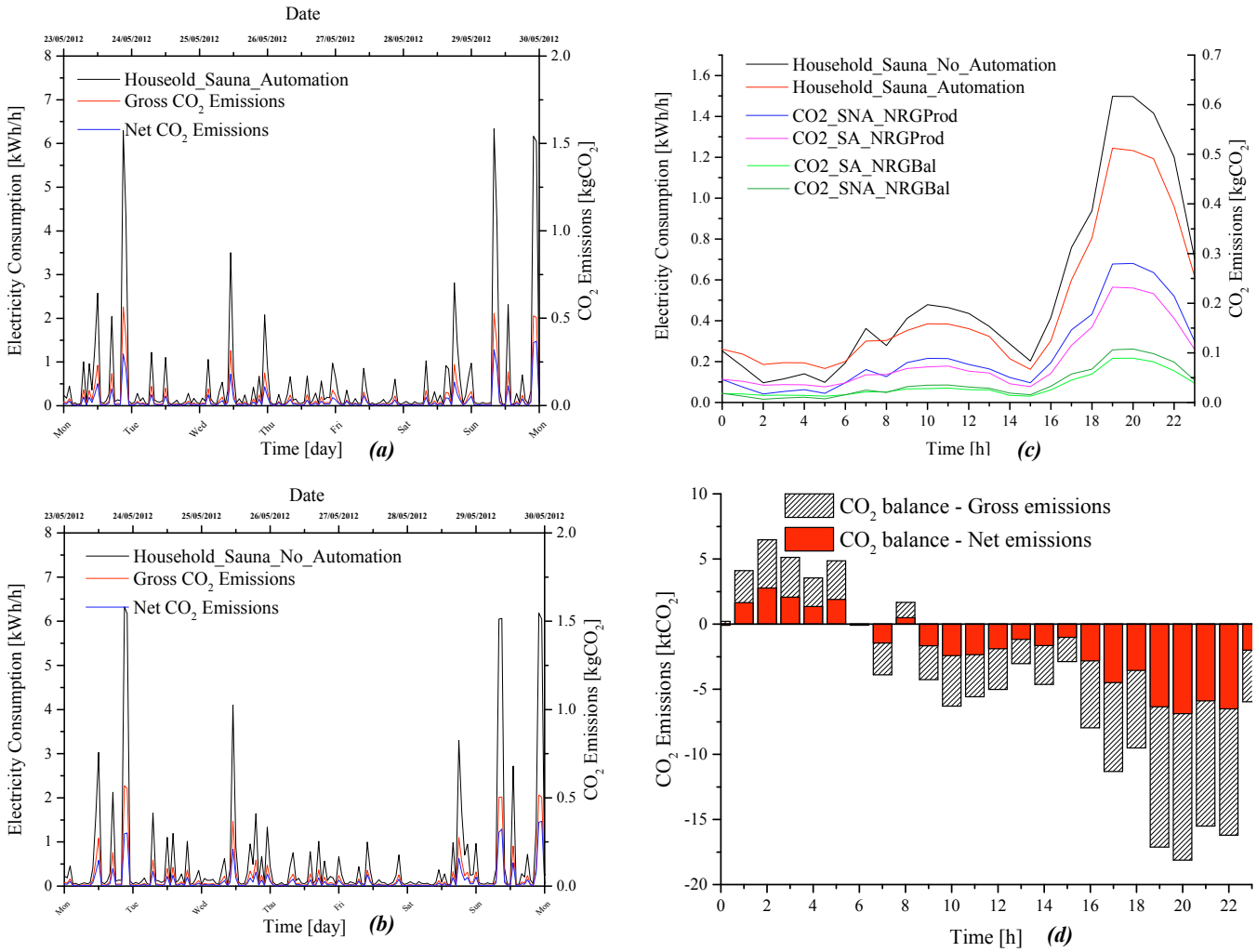


Figure 7. Electricity consumption with its related CO_2 emissions for (a) a dwelling with automation and (b) a dwelling without automation, (c) Daily electricity consumption profiles and (d) the related CO_2 emissions balance between both houses in terms of gross and net emissions

hourly basis using different components for evaluating the electricity consumption from appliances, without primary and secondary electric heating systems. Two dwellings were studied: one with home automation and one without home automation, and the difference in their CO_2 balance was evaluated.

The modelled house contains twenty-one appliances, all of them labelled A or B [14]. The house, being in Finland, has an electric sauna stove of 6 kW. The overall electricity consumption of appliances in this modelled house is 4 501 kWh/y, which correlates with the findings of the European ODYSSEE MURE project and that of the Sähkötohtori Analysis [25]. The measured data were obtained from detached houses in Oulu, Finland, which were equipped with a 10 kW sauna stove.

B. Impact of the simulated home automation on the emissions

The model showed that the CO_2 emissions are highly dependent on electricity consumption levels. Depending on the energy mix for electricity production at a given time, CO_2 emission levels may even be lower at peak hours and thus not proportional to consumption levels. Two models have been developed. In the first case, the CO_2 emissions from the house are accounted relatively to the electricity production only. In the second case, the CO_2 emissions are balanced with the electricity exported and imported. Fig. 7 represents the energy consumption for the two cases: with home automation (Fig. 7 (a)), and without home automation (Fig. 7 (b)). The electricity consumption shown was extracted for a randomly selected week in May 2012, starting on Monday, the 23rd of May.

1) Case 1: Emissions related to electricity production

The houses in the two cases are similar in their characteristics such as number and types of appliances, number of inhabitants, dimensions, users' habits. The CO_2 emission levels vary from 0.06 to 0.20 $kgCO_2/kWh$. The levels depend on the energy mix of Finland's electricity generation. Consequently, the hourly-based emissions peak at 1.93 $kgCO_2/h$ for the house without home automation and

1.81 $kgCO_2/h$ for the one with home automation. In the first case, the related energy demand was 10.03 kWh/h and in the second 9.42 kWh/h. The maximum electricity consumptions are 12.33 kWh/h, and 10.16 kWh/h. The emission peaks are somewhat related to the level of electricity consumption but also to the energy mix for electricity generation. The use of home automation may reduce the instantaneous peak of CO_2 emissions. The daily electricity profile of the houses and CO_2 balance between the two cases are represented in Fig. 7 (c). The difference in the profile of the two modelled houses result in a 592 kWh/y reduction of total electricity consumption. In terms of CO_2 emissions, the house that is not equipped with a home automation emits 543 $kgCO_2/y$, while the house with home automation emits 473 $kgCO_2/y$. The amount of CO_2 saved represents 12.78 % of original emissions.

Home automation shifted some of the electricity consumption from the evening peak to the night. It resulted in a decrease of CO_2 emissions in the evening down to 37 % from the original levels, and an increase of 51 % of CO_2 emissions overnight (Fig. 7 (d)). Considering, however, that the emissions overnight are about 0.1 $kgCO_2/h$ on average, this can be regarded as a relatively small, cumulative amount.

The emissions increased overnight by 3 to 5 $kgCO_2$, and reduced by 17 $kgCO_2$ on average over the whole year during the evening. While the home automation was not optimised for reducing CO_2 emissions but rather for cutting peak load consumption, it resulted in the decrease of CO_2 emissions as well. Notwithstanding, it is to be seen that the emissions related to electricity generation countrywide vary throughout the day. Fig. 8 represents the summed CO_2 emissions per hour on the left axis and the hourly average profile of CO_2 emissions on the right axis for the year 2012 from the electricity produced in Finland. The CO_2 emissions during the peak hours are 0.95 $ktCO_2/h$ on average, and add up to a total of 346 $ktCO_2$ between 6 and 7 pm. The lowest point on the daily plot of CO_2 emissions occurs around 2 and 3 am, with an average emission of 0.8 $ktCO_2/h$ and a corresponding emission for this hour throughout the year is 294 $ktCO_2$.

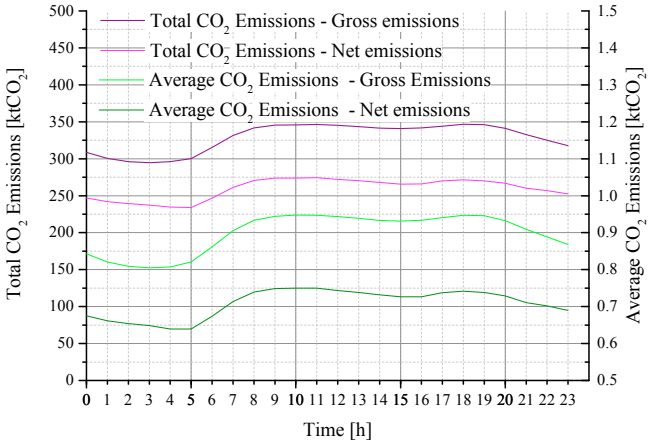


Figure 8. Total and average daily profile of the carbon dioxide emission in 2012, Finland

2) Case 2: Emissions related to net load

The CO_2 emissions in this second case were found much lower than in Case 1. Firstly, the total CO_2 emissions factor $E_{h,i\text{-}gen}$ has slightly decreased. This can be interpreted as an improvement in the net CO_2 emissions from electricity at the country level. This is explained by the fact that Finland is importing its electricity mostly from Sweden, and Sweden has an average emission factor around 7 times smaller than that of Finland. On the other hand, Finland is exporting electricity with a relatively high emission factor. As well, the emissions from Finnish electricity have been calculated for every hour and, therefore, there are peaks of CO_2 emissions. Conversely, the electricity from the neighbour countries are applied a constant factor, thus this bring a bias result. Nonetheless, the exchange of electricity is beneficial for Finland in terms of CO_2 emissions. In this case, the house had a 335 $kgCO_2/y$ emission without home automation, and 293 $kgCO_2/y$ with home automation. This means a difference of about 38 % between Case 1 and Case 2. This also indicates that CO_2 emissions can be interpreted very differently, depending on whether the emissions associated with exported electricity are subtracted from the total CO_2 emissions of the country or included in it. Similarly to Case 1, the peaks of CO_2 emissions are reduced, and are about 24 % lower than in Case 1. In Case 2, the CO_2 peak for the house without home automation is 1.46 $kgCO_2/h$, and 1.37 with home automation.

At the system level, the total and average hourly CO_2 emissions have decreased as well. In case the exported and imported electricity are accounted in net emissions, the low peak occurs between 4-5 am with an average emissions of 0.64 $ktCO_2$ and the high peak period occurs between 10-11 am with an average emissions of 0.75 $ktCO_2$ and cumulates to 275 $ktCO_2$ in same hour. Regarding the shift of CO_2 emissions due to the home automation device and the feedback strategies used for informing the private consumers, it has decreased by 6 $kgCO_2$ in the evening and has risen by 2.7 $kgCO_2$ in the night time. The quantities of CO_2 shifted, as presented in Fig. 6 (d), and are different from Case 1 and Case 2, as the CO_2 emission profiles for both cases are different (see Fig. 8).

VI. Discussion

The methodology developed in this article allows the integration of physical electricity production variation and resource usage. The emission factor is re-calculated every month for each technology, thus better reflecting the seasonal variations compared to a fixed emission factor. The gross emission factor can be calculated from 2004, but net emissions from electricity production are only available since 2012. This method has the advantage to use publicly available information with an hourly window grid. Nevertheless, the method presented is restricted to data availability from country-to-country. It is thus challenging to evaluate the replicability level of the method.

When studying the impact of home automation, both cases

TABLE III. CO_2 EMISSIONS SUMMARY FOR THE TWO STUDIED CASES

	CO_2 emissions relative to		
	Electricity produced (Gross Emissions)	Net electricity consumed (Net Emissions)	Unit
Min. $E_{h,i\text{-}gen}$	0.06	0.04	$kgCO_2/kWh$
Max. $E_{h,i\text{-}gen}$	0.20	0.19	
Max $E_{h,house}$ SA	1.81	1.37	$kgCO_2/h$
Max $E_{h,house}$ SNA	1.93	1.46	
Max $P_{i,house}$ SA	12.33	12.33	kWh/h
Max $P_{i,house}$ SNA	10.16	10.16	
Total $E_{h,house}$ SA	543	335	$kgCO_2/a$
Total $E_{h,house}$ SNA	473	293	
Max Average $E_{h,i\text{-}gen}$	0.95	0.75	$ktCO_2$
Min Average $E_{h,i\text{-}gen}$	0.8	0.64	
Max Sum $E_{h,i\text{-}gen}$	346	275	
Min Sum $E_{h,i\text{-}gen}$	294	234	

showed that load shifting can contribute to 12.7 % decrease in CO_2 emissions. However, there is a difference depending on whether the balance of import and export is considered. As well, consumer awareness and their willingness to comply is also a factor in the potential for reducing CO_2 emissions. Table III summarises the results from the CO_2 emissions and the electricity consumption from both houses. It is necessary to point out the importance of methods evaluating emissions on the results. It is paramount that the countries involved use the same methodology for their CO_2 evaluation. In this study, Finland is mostly importing electricity from Sweden and Russia and exporting to Norway and Estonia. For Sweden, this means importing "polluted" electricity and exporting cleaner electricity to Finland. Consequently, for Finland, the shifting of CO_2 emissions is greater when compared to the emissions of gross electricity production. It also needs to be pointed out that the house simulator can be used for either optimising electricity consumption or CO_2 emissions or both. In our case, the multi-objective algorithm was developed for optimising electricity consumption, but it also resulted in emission reductions. To optimise for CO_2 reduction, would be an additional challenge. In addition, an added level of complexity is whether export/import net emissions are considered or not.

VII. CONCLUSION AND FUTURE WORK

The article detailed the CO_2 emissions of electricity generation in Finland. Firstly, CO_2 emissions from electricity production and trade have been evaluated using a methodology developed within this research. Then, monthly, weekly, and daily data of electricity generation were used to calculate corresponding CO_2 emissions into hourly data. This was used to evaluate the CO_2 emission profile of households. The model was based on hourly electricity load profiles previously built. The methodology developed reflects the seasonal variations as well as the monthly fluctuation in resources usage from the power plants. It, in turn, increases the reliability for evaluating the CO_2 emissions due to the electricity consumption.

Secondly, the CO_2 emissions associated with imported and exported electricity generation were accounted as well. Both cases show the same peak distribution in their daily profile. Notwithstanding, emissions will depend on the fuel used at a particular hour. Therefore, the relationship between electricity production, import and export is not straightforward. The cumulated CO_2 emissions overnight from the electricity produced in Finland stand at around 290 $ktCO_2/h$, while the peak reaches 345 $ktCO_2/h$. Considering the import and export of electricity, and their related CO_2 emissions, the peak dropped to 230 $ktCO_2/h$ overnight, and the high peak is at 275 $ktCO_2/h$.

Although the home automation was not optimised for emission reduction, the CO_2 emissions are somewhat proportional to electricity consumption levels. The study showed that home automation might reduce the carbon dioxide emission by 12.7 % while influencing the private consumers' everyday routine. The CO_2 emissions have been reduced most substantially during the evening peak, by 18 $kgCO_2/h.y^{-1}$ in the first case and by 6 $kgCO_2/h.y^{-1}$ in the second case, while the emissions at night time have increased from 3 to 5 $kgCO_2/h.y^{-1}$ on average. Although the CO_2 emissions related to electricity consumption from appliances are strongly correlated, the energy mix for producing this electricity needs to be considered and thus optimised for reducing the carbon footprint of households.

Consequently, smart buildings within a smart grid may not only participate in load shifting and increase energy efficiency or decrease electricity consumption, but they can also significantly contribute to the reduction of CO_2 emissions. It will, in turn, impact the total CO_2 emissions of the country and will assist in achieving the decarbonisation goal of the EU.

The limitation of this research is that there was no information available on the variation of the energy mix from exporting countries and, therefore, import electricity had to be considered with a yearly constant CO_2 emission factor. Secondly, in the case of Finland, a more detailed estimation would require knowing the energy mix hour-by-hour, rather than estimating it from the monthly average. However, currently, this information is not available in Finland.

Further research will investigate the impact of private consumers in correlation with home automation for reducing the CO_2 emissions of households. In addition, a full assessment considering district-heating systems ought to be done, in order to achieve full integration of smart buildings in a SEN. Finally, the multi-objective algorithms will have to be further developed and improved on.

Acknowledgement

The Thule Institute Doctoral Programme is acknowledged for financing this research.

References

[1] J.-N. Louis, A. Caló, and E. Pongrácz, "Smart houses for energy efficiency and carbon dioxide emission reduction," ENERGY 2014 : The Fourth International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies, pp. 44–50, Apr. 2014.

[2] L. C. De Silva, C. Morikawa, and I. M. Petra, "State of the art of smart homes," Engineering Applications of Artificial Intelligence, vol. 25, no. 7, pp. 1313–1321, Oct. 2012.

[3] S. J. Darby, J. Stromback, and M. Wilks, "Potential carbon impacts of smart grid development in six European countries," Energy Efficiency, vol. 6, no. 4, pp. 725–739, May 2013.

[4] Publications Office, Directive 2012/27/EU of the European Parliament and of the Council of 25 October 2012 on energy efficiency, amending Directives 2009/125/EC and 2010/30/EU and repealing Directives 2004/8/EC and 2006/32/EC Text with EEA relevance. 2012, pp. 1–56.

[5] TEM, Valtioneuvoston asetus sähköntoimitusten selvityksestä ja mittauksesta. 2013, pp. 1–6.

[6] D. Long Ha, S. Ploix, M. Jacomino, and M. Hoang Le, "Home energy management problem: towards an optimal and robust solution," in Energy Management, no. 5, F. Macia Perez, Ed. InTech, 2010, pp. 77–106.

[7] H. Hens, G. Verbeeck, and B. Verdonck, "Impact of energy efficiency measures on the CO2 emissions in the residential sector, a large scale analysis," Energy & Buildings, vol. 33, pp. 275–281, Jan. 2001.

[8] N. Pardo and C. Thiel, "Evaluation of several measures to improve the energy efficiency and CO2 emission in the European single-family houses," Energy & Buildings, vol. 49, pp. 619–630, Jun. 2012.

[9] A. de Almeida, P. Fonseca, B. Schlomann, N. Feilberg, and C. Ferreira, "Residential monitoring to decrease energy use and carbon emissions in europe," presented at the EEDAL'06, 2006, pp. 1–17.

[10] V. A. Dakwale, R. V. Ralegaonkar, and S. Mandavgane, "Improving environmental performance of building through increased energy efficiency: A review," Sustainable Cities and Society, vol. 1, no. 4, pp. 211–218, Dec. 2011.

[11] N. Gilbraith and S. E. Powers, "Residential demand response reduces air pollutant emissions on peak electricity demand days in New York City," Energy Policy, vol. 59, no. C, pp. 459–469, Aug. 2013.

[12] M. Stokes, "Removing barriers to embedded generation : a fine-grained load model to support low voltage network performance analysis," Institute of Energy and Sustainable Development de Montfort University, Leicester, 2005.

[13] A. Grandjean, J. Adnot, and G. Binet, "A review and an analysis of the residential electric load curve models," Renewable and Sustainable Energy Reviews, vol. 16, no. 9, pp. 6539–6565, Dec. 2012.

[14] J.-N. Louis, A. Caló, K. Leiviskä, and E. Pongrácz, "Home automation for a sustainable living – Modelling a detached house in Northern Finland," Proceedings of the 7th International Conference on Energy Efficiency in Domestic Appliances and Lighting EEDAL'13, pp. 561–571, Sep. 2013.

[15] N. Arghira, L. Hawarah, S. Ploix, and M. Jacomino, "Prediction of appliances energy use in smart homes," Energy, vol. 48, no. 1, pp. 128–134, Dec. 2012.

[16] R. Missaoui, H. Joumaa, S. Ploix, and S. Bacha, "Managing energy smart homes according to energy prices: analysis of a building energy management system," Energy & Buildings, vol. 71, pp. 155–167, Mar. 2014.

[17] P. Chavali, P. Yang, and A. Nehorai, "A distributed algorithm of appliance scheduling for home energy management system," IEEE Trans. Smart Grid, vol. 5, no. 1, pp. 282–290, Jan. 2014.

[18] R. Huang, M. Itou, T. Tamura, and J. Ma, "Agents based approach for smart eco-home environments," presented at the The 2010 International Joint Conference on Neural Networks (IJCNN), 2010, pp. 1–8.

[19] P. Stoll, N. Brandt, and L. Nordström, "Including dynamic CO2 intensity with demand response," Energy Policy, vol. 65, no. C, pp. 490–500, Feb. 2014.

[20] IEA, International Energy Agency, "CO2 emissions from fuel combustion highlights 2013," pp. 1–158, Oct. 2013.

[21] Official Statistics Finland, Ed., Energy Consumption in households by energy source in 2012, 2012. [Online]. Available: http://www.stat.fi/til/asen/2012/asen_2012_2013-11-13_tau_002_en.html. [Accessed: 05-Dec-2013].

[22] E. Tetri, A. Sarvaranta, and S. Syri, "Potential of new lighting technologies in reducing household lighting energy use and CO2 emissions in Finland," Energy Efficiency, pp. 1–12, Dec. 2013.

[23] M. G. Ippolito, E. R. Sanseverino, and G. Zizzo, "Impact of building automation control systems and technical building management systems on the energy performance class of residential buildings: An Italian case study," Energy & Buildings, vol. 69, pp. 33–40, Feb. 2014.

[24] Official Statistics Finland, Ed., Household-Dwelling units by number of persons 1960 - 2012. [Online]. Available: http://www.stat.fi/til/asas/2012/asas_2012_2013-05-22_tau_001_en.html. [Accessed: 05-Dec-2013].

[25] V. Rouhiainen, "Decomposing electricity use of Finnish households to appliance categories," presented at the Energy efficiency in domestic appliances and lighting , Proceedings of the 5th international conference EEDAL '09, Berlin, 2010, vol. 2, pp. 438–455.

[26] Finnish Industry Association, "Power generation in Finland - fuels and CO2-emissions," Energiateollisuus, Helsinki, Nov. 2013.

[27] Finnish Industry Association, "Monthly electricity supply in Finland May 2014," Finnish Industry Association, Jun. 2014.

[28] S. Heier, Grid Integration of Wind Energy, 3rd Edition. John Wiley & Sons Ltd, 2014, pp. 1–520.

[29] V. Turkia and H. Holttinen, "Wind energy statistics in Finland 2013," VTT Technical Research Centre of Finland, Jun. 2014.

[30] Fingrid Oyj, Ed., Load and generation. [Online]. Available: http://www.fingrid.fi/en/electricity-market/load-and-generation/Pages/default.aspx. [Accessed: 04-Dec-2013].

[31] Finnish Industry Association, "Electricity net production, imports and exports, in Finland," no. 258. Energiateollisuus, Helsinki, 28-Nov-2013.

[32] Energy Authority, "Energiaviraston voimalaitosrekisteri," no. 172. Energy Authority, Helsinki, 04-Jan-2014.

[33] Energiamarkkinavirasto, "Emission Allowance Balance." EMV, Helsinki, 28-Nov-2013.

[34] J.-N. Louis, "Smart Building to improve energy efficiency in the residential sector," Oulu University, 2012.

Incorporating Reputation Information into Decision-Making Processes in Markets of Composed Services

Alexander Jungmann
C-LAB
University of Paderborn
Paderborn, Germany
Email: alexander.jungmann@c-lab.de

Ronald Petrlic
CISPA
Saarland University
Saarbrücken, Germany
Email: ronald.petrlic@uni-saarland.de

Sonja Brangewitz
Department of Economics
University of Paderborn
Paderborn, Germany
Email: sonja.brangewitz@wiwi.upb.de

Marie Christin Platenius
Heinz Nixdorf Institute
University of Paderborn
Paderborn, Germany
Email: m.platenius@upb.de

***Abstract*—One goal of service-oriented computing is to realize future markets of composed services. In such markets, service providers offer services that can be flexibly combined with each other. However, although crucial for decision-making, market participants are usually not able to individually estimate the quality of traded services in advance. To overcome this problem, we present a conceptual design for a reputation system that collects and processes user feedback on transactions, and provides this information as a signal for quality to participants in the market. Based on our proposed concept, we describe the incorporation of reputation information into distinct decision-making processes that are crucial in such service markets. In this context, we present a fuzzy service matching approach that takes reputation information into account. Furthermore, we introduce an adaptive service composition approach, and investigate the impact of exchanging immediate user feedback by reputation information. Last but not least, we describe the importance of reputation information for economic decisions of different market participants. The overall output of this paper is a comprehensive view on managing and exploiting reputation information in markets of composed services using the example of On-The-Fly Computing.**

Keywords–Reputation, Service Market Interactions, Economic Decisions, Service Composition, Service Matching, Learning Service Recommendation, Game Theory.

I. PREFACE

This paper is a revised and expanded version of our paper entitled 'Towards a Flexible and Privacy-Preserving Reputation System for Markets of Composed Services' (SERVICE COMPUTATION) [1]. In addition to the original paper's scope, this expanded version covers new research results regarding the incorporation of reputation information into processes that are crucial in markets of composed services: fuzzy service matching, service composition including a learning service recommendation component, and economic decisions with respect to service market interactions. Moreover, the requirements imposed on the reputation system are expanded to account for context-specific reputation and expert ratings.

II. INTRODUCTION

A major goal of On-The-Fly (OTF) Computing is the automated composition of software services that are traded in dynamic markets and that can be flexibly combined with each other [2][3]. A user formulates a request for an individual software solution, receives an answer in terms of a composed service, and finally executes the composed service.

As an illustrative example, let us assume that someone wants to post-process a holiday video. However, it does not pay off to use a monolithic software solution because such software provides a lot of dispensable functionality, and is therefore too expensive to buy for just this purpose. What this person needs is an individually customized software composed of only those services, which together are able to satisfy his needs. A famous web-based application for individual post-processing tasks is Instagram, which provides different image processing services that can be applied to an uploaded photo or video [4]. However, the variety of available services is restricted and the selection of appropriate services has still to be done manually.

Now, let us consider a market of image processing services. A person, who wants to post-process his video, becomes a user within this market by formulating a request describing what he expects from the composed service (e.g., the functionality to create videos with reduced image noise and an increased brilliance homogeneously distributed throughout the entire video). Subsequently, a post-processing solution that satisfies the user's request is automatically composed based on image processing services that are supplied by different market participants. In this scenario, the user only has to pay for the actually utilized functionality. Figure 1 shows an exemplary use case in terms of a single photo for such a market of image processing services. The original photo depicted in Figure 1a was captured and shall be post-processed in order to change the appearance according to individual user preferences. To achieve the different effects shown in Figure 1b to Figure 1d, the functionality of a single service is usually not sufficient. In fact, composed services have to be constructed based on image processing services that are available in the market. In

(a) (b) (c) (d)

Figure 1. Original image (a) is processed by different composed services to achieve the different effects shown in (b), (c), and (d).

order to compose services that satisfy specific requirements, the quality of services regarding functional as well as non-functional properties must be taken into account.

However, for market participants who are willing to buy those services, it is difficult to estimate the quality of services before the services are actually used. For example, an image processing service's response time can be predicted to a certain extent, but it is very dependent on the specific context, e.g., its execution environment and its current load. Other markets such as eBay or Amazon solve this problem by using a reputation system. Within such a system, the experiences other users made in previous transactions are collected. Thereby, the reputation information provides new users an indicator for the service quality they can expect. As an example, let us consider that many users were entirely satisfied with a specific image processing service and rated it with five stars, for example. As a consequence, this service gained a high reputation, which makes it more attractive for future users. Not only the requesters, but also the whole market benefits from considering reputation, because the providers of high-quality products are rewarded with a high reputation, thereby increasing their chances for future sales. On the other hand, low-quality or even deceptive service providers will only be able to sell their services at low prices or will vanish from the market after some time, which again pays off for all customers. Existing reputation systems used by eBay or Amazon, for example, do not explicitly consider ratings for composed services. Other reputation systems, such as those to rate trips or hotels, often ask the user to evaluate different aspects. However, single services cannot be combined with each other as flexibly as needed in the OTF Computing market. Thus, a reputation system for composed services is still an open challenge.

The contribution of our original paper covers the identification of requirements for a reputation system for markets of composed services such as OTF Computing [1]. Furthermore, it covers the conceptual design of our proposed solution in terms of a flexible reputation system. The extended contribution of this revised version additionally covers, besides an expanded list of requirements, a more detailed description of selected processes that are essential for realizing markets of composed services. Each of these processes incorporates reputation information according to our proposed reputation system. Technical details of a prototypical implementation of our reputation system, however, are not part of the contribution and are consequently beyond the scope of this paper. The contribution of this paper is not necessarily restricted to OTF Computing alone. Results of our work can also be adopted to other areas, in which reputation of combinable products is vital.

To the best of our knowledge, there are currently no existing reputation system approaches that can be directly applied to markets of composed services like, e.g., the OTF Computing market. Reputation has already been considered in the area of service composition: A survey is presented by Gómez Mármol et al. [5]. However, each of the existing approaches only deals with a subset of the requirements we identified. For example, context-specific reputation and the consideration of expert ratings is missing in almost all related approaches and often only the reputation of services but no reputation of other market participants is considered. Furthermore, privacy protection is not considered by already existing approaches. Reputation systems that take privacy protection into account explicitly, either entail a high overhead, or privacy is only a "property", which is said to be achieved—but not enforced cryptographically.

This paper is organized as follows. Section III introduces OTF Computing while mainly focusing on those aspects that are relevant for the work at hand. Furthermore, it motivates the significance of reputation in OTF Computing. Section IV gives a detailed problem description by subsequently introducing crucial requirements for a reputation system in OTF Computing. Section V presents our conceptional solution in terms of a flexible reputation system that covers all identified requirements. Existing approaches that only partially cover these requirements are discussed in Section VI. Section VII introduces our fuzzy service matching approach that focuses on reputation information about services. Section VIII, in turn, mainly deals with the incorporation of reputation information about composed services into our composition approach. Economic decisions with respect to service market interactions under consideration of available reputation information are investigated in Section IX. Based on the heretofore presented results, Section X describes remaining research challenges that will be addressed in our future work. Finally, the paper concludes with Section XI.

III. On-The-Fly Computing

A major goal of OTF Computing is automated composition of flexibly combinable services that are traded in markets. A user's request for an individual software solution should be resolved by automatically composing a solution on demand. OTF Computing addresses the entire process, starting with fundamental concepts for organizing large-scale service markets up to the final execution of a composed service. Embedding automatic service composition into service markets is one key challenge for realizing OTF Computing. The whole OTF Computing process is very complex, as a number of different aspects need to be taken account of in an integrated fashion. In this paper, we cover three of those aspects exemplarily. At that time, we do not have an implementation of the whole OTF Computing scenario. We do have prototype implementations of the individual processes, though, which should be combined in the future in order to be able to prove the practicality of OTF Computing as a whole.

A. Automatic Service Composition

In general, we interpret automatic service composition as the sequential application of composition steps. A composition step may, for example, correspond to selecting a service in order to realize a placeholder within a workflow [6]. Regarding our initial example in terms of image processing services, a placeholder could correspond to a class of services, which provide similar functionality (such as smoothing filters). For execution, a specific service (e.g., Gaussian smoothing) must then be selected. A composition step, however, may also correspond to a single step within a composition algorithm based on Artificial Intelligence (AI) planning approaches [7][8][9][10].

For simplicity, let us assume that a workflow is available and that a service composition step corresponds to selecting a service. We divide a single composition step into two separate processes, which subsequently reduce the amount of qualified service candidates. First of all, a *Service Matching* process determines to what extent a particular service fulfills a placeholder's functional (e.g., signatures and behavior) as well as non-functional requirements (e.g., quality properties such as response time or reliability) [11][12]. Based on the matching result, services that provide significantly different functionality or that violate important non-functional restrictions can be discarded directly. Subsequent to the matching process, a *Service Recommendation* process identifies (and ranks) the best service candidate(s) out of the set of remaining services. During the recommendation process, explicitly given non-functional objectives regarding the final composed service (e.g., maximizing the performance while simultaneously minimizing the costs) as well as implicit knowledge from previous composition processes (e.g., a certain service is more qualified in a particular context than others) are incorporated. The incorporation of knowledge from previous composition processes is realized by means of Reinforcement Learning (RL) [13], and requires feedback about the quality of the execution result [14].

B. Market Infrastructure Perspective

Figure 2 shows the transactional view on the entire OTF Computing process, reduced to those processes that are relevant for the work at hand. *OTF Provider Selection* and *Service Provider Selection* are decision-making processes regarding transactions within the market. Three different classes of market participants are involved in the overall process: users, OTF providers, and service providers. A user formulates a request for an individual software solution and sends it to an OTF provider of his choice (*Step 1*). The selected OTF provider processes the request and automatically composes a solution based on elementary services that are supplied by independent service providers.

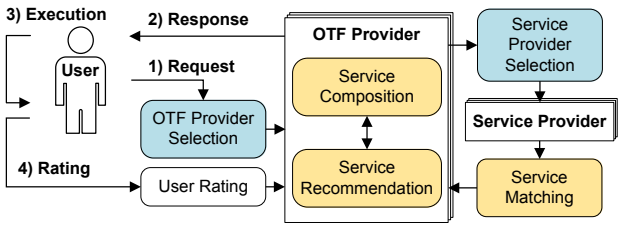


Figure 2. Overall OTF Computing process.

For each composition step, an OTF provider asks a selected subset of service providers for elementary services. The previously mentioned matching process is part of the OTF architecture and takes place before an OTF provider receives answers about appropriate elementary services. The matching process operates as a filter ensuring that only services that fulfill the desired requirements to a certain extent are returned. The recommendation process, in turn, is part of the OTF provider-specific composition process and highly depends on the context of the request.

As soon as a composed service is created, it is passed on to the user (*Step 2*), who subsequently executes it (*Step 3*). After execution, the user rates his degree of satisfaction regarding the quality of the execution result (*Step 4*). In the current setting, the value of the user rating is immediately returned to the associated OTF provider. By transforming the value into a reward and incorporating it into the RL process within the recommendation system, the OTF provider improves its internal composition strategy (recommendation process) for future user requests [15].

C. Reputation as a Signal for Quality

In a dynamic market of software services, information about quality (e.g., service quality or the quality of OTF providers) is essential. A user may resort only to OTF providers of a certain quality (e.g., with respect to customer support), while simultaneously accepting only composed services of a certain quality level (e.g., composed services with high reliability and trustworthiness). OTF providers, in turn, have to build composed services consisting of elementary services with a quality level according to a user's request. Information about quality, however, is either difficult to estimate before a transaction actually took place, or cannot be simply trusted if the quality information is provided by the associated market participant itself (e.g., when a service provider specifies the quality of his own services). Our solution to overcome these issues is to replace the previously mentioned and fairly simple user rating procedure (cf. Figure 2) with a flexible reputation system, which aggregates user ratings into single reputation values and provides them to market participants. Reputation can then be incorporated as an estimation of quality into the different decision-making processes.

IV. Problem Description and Requirements

Our goal is to explicitly incorporate reputation information as an estimation of quality into the OTF Computing process. Using goal-oriented requirements engineering [16], we systematize our reputation system requirements by investigating the role of reputation from different perspectives.

A. Reputation Information Within the OTF Process

As shown in Figure 2, the OTF Computing process is initiated by a user's request. To enable users to choose an OTF provider they want to establish a business relationship with, i.e., to buy a composed service from, reputation information about OTF providers must be available.

(*R1*) *OTF Provider Reputation:* The reputation system must provide reputation information about OTF providers.

The selected OTF provider has to ensure that the requested composed service satisfies the user's requirements regarding reputation. For this purpose, the reputation of service providers and the reputation of their supplied elementary services has to be considered during the composition process. In order to enable OTF providers to select service providers they want to retrieve elementary services from, reputation information about service providers must be available.

(*R2*) *Service Provider Reputation:* The reputation system must provide reputation information about service providers.

Reputation of elementary services influences the reputation of composed services. For example, if a composed image processing service uses a well-known, reputable implementation of a specific image filter, it can be assumed, that the composed service's reputation will be higher than the reputation of a service composed of unknown elementary services. Thus, the service matching processes (cf. Figure 2) as well as the service recommendation process have to consider the reputation of elementary services. While the matching process has to determine to what extent an elementary service fulfills certain requirements considering reputation, the recommendation process has to determine the best composition steps including reputation. Reputation information, however, cannot be simply extrapolated from service providers to elementary services, since a service provider may supply services of varying quality. Therefore, reputation information about elementary services must be available, too.

(*R3*) *Service Reputation:* The reputation system must provide reputation information about elementary services that have been consumed as a part of a composed service.

The recommendation process additionally rates alternative composition steps based on experience gained from previous composition processes. Reputation information about previously composed services is needed as feedback for the recommendation process in order to adapt its recommendation strategy by means of RL. An OTF provider's experience, however, can be considered a business secret that must not be revealed to other market participants.

(*R4*) *Composed Service Reputation:* The reputation system must provide reputation information about composed services without revealing business secrets of OTF providers.

Up to now, we focused on the overall reputation of (composed) services and providers. However, in reality, reputation is rather context-specific [17]. For example, an image processing service could have a good reputation regarding the response time but a bad reputation regarding security. Thus, the reputation system should maintain vectors instead of single values for ratings and reputation. This provides requesters with the possibility to specify more detailed requests, e.g., "I want an image processing service with a high reputation with respect to security, but its reputation for response time is not that important to me".

(*R5*) *Context-specific Reputation:* The reputation system must distinguish between reputation values based on different contexts, i.e., based on different properties of the rated service or provider.

Users only interact with OTF providers and not with service providers directly (cf. Figure 2). As a consequence, a user's ratings mainly contain information about OTF providers and their composed services. Only once in a while may a user be able to additionally rate elementary services. For example, when using a composed service for an image processing task, users may not be aware of all elementary services, e.g., of the filter service that reduces image noise. However, they may be able to rate an elementary service that implements an image compression algorithm, since the way the algorithm effects the execution result can be directly observed in terms of the size and quality of the generated image or video.

(*R6*) *Incomplete User Rating:* The reputation system has to consider that a user is most often just able to rate OTF providers and their composed services, while a user is only sometimes able to rate elementary services and never able to rate service providers.

In certain scenarios, users might be especially interested in ratings by *experts*, i.e., people who have a well-known expertise in that domain. The reputation system should allow for a flexibility when it comes to dealing with ratings provided by experts. As an example, the system could support weighting experts' ratings more than ratings provided by non-experts, i.e., "ordinary" users.

(*R7*) *Expert Ratings:* The reputation system shall provide a flexible mechanism of handling experts' ratings.

B. Technical Requirements

The reputation system needs to provide access to the different reputation values mentioned in the previous section for the different parties illustrated in Figure 2. Those parties have diverse and variable needs for reputation value computations and access as well as interaction preferences. For the service recommendation process, recent ratings are more important to accelerate the learning process and, therefore, reputation value computations that put a higher weight on those ratings are desired (e.g., rather a geometric mean than an average with equal weights). In contrast, for a user, it might be preferable that a certain composed service has a very low failure rate and, thus, during the provider selection process, reputation values that include historic values to a sufficient extent and put a higher weight on negative ratings have to be considered. The reputation system's functionality to process user ratings and to provide them as reputation information has to satisfy the diverse needs of the requesting parties.

(*R8*) *Flexible Processing of User Ratings:* The reputation system must support flexible processing of user ratings.

Certain restrictions may be applied: Concerning requirement (*R4*), reputation information about composed services shall be

retrievable only by the OTF provider that originally accomplished the service composition process.

(*R9*) *Access Control:* The reputation system must implement access control to reputation values.

Furthermore, the reputation system shall support different interaction models. Parties, such as the OTF provider's service recommendation component, need new reputation information as soon as it is available. New reputation information has to be automatically forwarded by the reputation system without explicitly asking for it. Other processes that rarely need to retrieve reputation information, such as users or the service matching component, shall be able to access those data actively on demand to reduce the data traffic.

(*R10*) *Interaction:* The reputation system must support alternative interaction concepts. Reputation information must either be provided on demand triggered by a request event, or actively sent to a party as soon as new reputation information is available.

Furthermore, security and privacy protection are crucial issues—as we have already investigated more generally for the OTF Computing as well [2]. If users could arbitrarily rate any services (without having used them), the reputation system would not constitute any benefit. If any party would be able to manipulate the reputation values, users could not trust the provided values and, thus, the reputation system's benefit would be lost as well.

(*R11*) *Rating Authorization:* Only authorized users, i.e., users that performed a transaction with an OTF provider, are allowed to rate that transaction. If a rating shall count as an "expert rating" (compare (*R7*)), a special authorization is needed: The expert needs to be recognized and authenticated as an expert during the rating process.

(*R12*) *Correctness:* The computed reputation value provided by the reputation system must be correct, i.e., it must not be possible for any party to manipulate the reputation value (computation).

Depending on the traded services in the market, users might only be willing to rate transactions if they can stay anonymous. They do not want to (publicly) reveal, which services were consumed by them. It has been shown in the past that designing a reputation system that provides user anonymity is a challenging task [18].

(*R13*) *Anonymity of Rating User:* No party shall be able to relate (individual) ratings to users. Even expert raters shall get the possibility to stay anonymous—if they want to; still, they need to be authenticated as experts (in order for their ratings to count more, for example) but their ratings shall not be linkable to them.

(*R14*) *Unlinkability of User Rating to Transaction:* The OTF provider must not be able to relate a rating to a transaction (previously executed with a certain user)—in order to achieve user anonymity.

V. A Flexible Reputation System

This section introduces the conceptual design of our proposed solution in terms of a flexible reputation system. First, the system's internal processes as well as its interaction capabilities are described. Afterward, we illustrate in particular how the system meets each requirement listed in Section IV. An overview of our proposed solution is given in Figure 3. It shows the internal structure of our flexible reputation system as well as the interactions with the OTF Computing process.

A. Basic Internal Structure

The reputation system is modeled as a stand-alone and independent component within the OTF Computing environment. The reputation values are derived by processing user ratings of services, composed services, as well as OTF providers. The internal structure can be divided into three main sections.

The *Accumulated Ratings* section provides functionality for accumulating raw values of incoming user ratings over time. To increase robustness, these values can be stored by means of a distributed storage system. The number of values to be stored is not necessarily restricted. However, depending on the available storage space and the amount of incoming values, outdated values may either be discarded or at least consolidated into a lower amount of values in the long run.

The *Aggregation System* provides functionality for processing a set of raw values in order to generate an aggregated representation. However, one can flexibly choose the set of raw values to be incorporated into the process, the actual aggregation function to be applied (e.g., arithmetic/geometric averaging, identifying the maximum or approximating the future trend by time series analysis) and the final representation (e.g., single scalars such as mean or median, or density functions in terms of their statistical parameters).

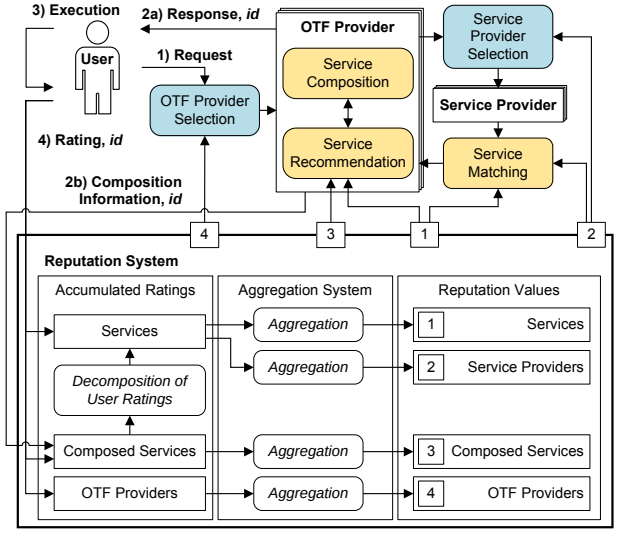


Figure 3. Proposed OTF Computing Reputation System. Internal structure and interactions with the OTF Computing process are depicted.

The *Reputation Values* section finally provides the interfaces for accessing the different reputation values of services, service providers, composed services, and OTF providers. When accessing reputation values, the set of raw user ratings to be considered, the actual aggregation function, as well as the final representation can be flexibly specified. Reputation values are not stored within the system, but always computed on demand dependent on the previously mentioned specifications. This flexibility allows requests for reputation information to adapt to more complex reputation requirements imposed by users. For example, a user may want an image processing service with a reputation value higher than 4 based on at least 20 user ratings that are not older than 6 months. Another user may want an image processing service, which has an average reputation value of 4, while no elementary service should have a reputation value less than 2.

B. Integration into the OTF Computing Process

Reputation values are consumed by the *Service Matching*, the *Service Recommendation*, the *Service Provider Selection*, and the *OTF Provider Selection* processes within the overall OTF Computing process. Besides flexibility regarding how a reputation value is internally computed, our proposed reputation system also provides flexible interaction capabilities. On the one hand, reputation values can be accessed by a *pull* approach whenever they are needed. Following this approach, the requester inherits the active role by asking for reputation data if and only if it is necessary. This solution is efficient when reputation information is needed less frequently (e.g., when a user wants to choose an OTF provider). On the other hand, a *push* approach shifts the active role to the reputation system. Reputation information is sent to a party as soon as new data is available. This approach also allows for creating a local cache of the latest reputation values without flooding the reputation system with redundant requests for possibly new information.

Figure 4 shows the interaction with the proposed reputation system using the example of the service matching process (matcher). During the OTF Computing process, the matcher is called for each elementary service that possibly satisfies an OTF provider's request (cf. Section III-B). In this context, Figure 4 illustrates the access of reputation information for exactly one elementary service by a *pull* approach.

The reputation matching process is initiated by providing the request information and the description of an elementary service and by calling the *match* operation. For the sake of simplicity, the request in the depicted example only shows an extract: An image processing service should have a minimum reputation value of 4. This request shall now be matched against an elementary service with id `ImagePro1`. The matcher asks the reputation system for a reputation value of service `ImagePro1` aggregated by means of an aggregation function with id `f_id`. Hence, the aggregation system fetches the relevant user rating values (3,4,3,5,5) from the storage, selects the corresponding aggregation function (here, arithmetic averaging), and computes an aggregated reputation value of 4. Based on this result, the matcher decides that the service matches to the request. We describe the challenges of reputation matching and our matching approach in more detail in Section VII.

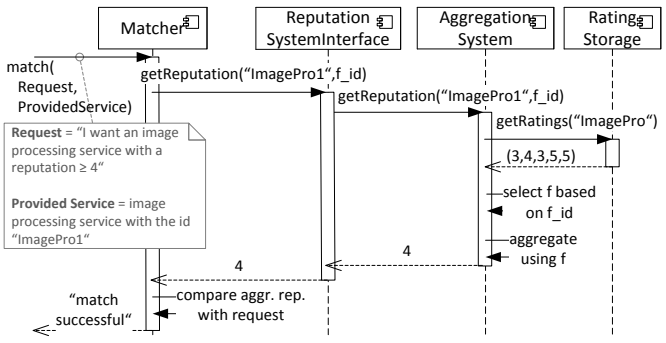


Figure 4. Simplified example interaction with the reputation system.

After a composed service was executed (*Step 3* in Figure 3), users are encouraged to provide feedback on their transactions. They are asked to rate composed services, OTF providers, and single services. The feedback in terms of user ratings is the foundation for generating reputation information within the reputation system. If the rating is provided by an expert, this information needs to be stored as well, as expert ratings can be given more weight than ordinary ratings—this depends on the requirements of the scenario and can flexibly be implemented as part of the reputation value computation as discussed before. To be able to identify, which composed service a rating belongs to, OTF providers attach an *id* to their response (*Step 2a* in Figure 3). This *id* corresponds to the particular structure of a composed service, meaning that identical composed services have identical *id*s. During the rating process for a composed service, this *id* is forwarded to the reputation system (*Step 4* in Figure 3).

Elementary services that are consumed as part of a composed service cannot always be rated separately by the user. In fact, due to complex user requests, we expect that this is rarely possible. Thus, in order to still be able to provide reputation values for elementary services and to benefit from all information available, our reputation system decomposes user ratings of composed services. To enable this decomposition, the *id* the OTF provider sends with his response (cf. Figure 3) is reused: Simultaneously with his response to the user (*Step 2a* in Figure 3), the OTF provider sends the same *id* together with *composition information* to the reputation system (*Step 2b* in Figure 3).

As pointed out above, our reputation system for OTF Computing shall provide flexibility, which also means that different implementations for the components are supported. We have already shown that such an implementation of a reputation system for the OTF Computing can be done in a *secure* and *privacy-preserving* way—respecting the requirements stated in Section IV [19]. In contrast to related work, as covered in Section VI, this approach only requires a single *reputation provider*, which is in line with the requirements of OTF Computing, and does not need any other components (such as a bulletin board). The approach is based on the Paillier cryptosystem [20] to provide a reputation value as an aggregation of individual user ratings without revealing anything about the individual ratings to any party. At the moment we are investigating if expert ratings can also be implemented in a privacy-respecting way. For that purpose,

we are investigating *group signatures* in more detail, as that mechanism shall allow the authentication of experts without revealing their identities.

C. Satisfying OTF Computing Requirements

Our proposed solution in terms of a flexible reputation system fulfills all requirements listed in Section IV. This section points out how the reputation system fulfills each of these requirements in particular.

The proposed reputation system enables users to rate OTF providers, composed services, and—if possible—elementary services. Assured by the transferred *id*, in this context, only users that are involved in a particular transaction taking place in the OTF Computing market, i.e., users that have requested, received and executed a particular composed service, are allowed to participate in the rating process. This ensures ratings by authorized users (*R11*). How to realize the rating process in particular (i.e., what kind of questions have to be asked and how a user rating value is represented) is beyond the scope of this paper.

Correctness of the provided reputation values is ensured by design. Reputation values are computed on demand by the system itself based on a pre-defined set of aggregation functions. Furthermore, the entire system is an independent component within the OTF Computing environment. As a consequence, manipulations of the computation process by other participants are eliminated (*R12*).

Anonymity of users (*R13*) as well as unlinkability of user ratings to transactions (*R14*) is ensured by the accumulation and aggregation functionality. For reasons of privacy protection, i.e., in order to not reveal individual user ratings, the reputation system always collects individual ratings and aggregates them. Although the single user ratings are stored within the reputation system, they are not accessible to market participants so that individual ratings are not traceable. In this context, it is important that the amount of accumulated user ratings is high enough and that the aggregation operation sufficiently condenses the user ratings such that it can be guaranteed that no information on individual ratings can be recovered. If not enough user ratings are included in the aggregation process (e.g., when not enough user ratings are available yet, or if a request explicitly specifies to only consider just a few user ratings), the reputation system will not provide a value but will raise an exception.

All processes that need reputation information within the entire OTF Computing process have access to the reputation system. The flexibility of our proposed solution enables each market participant to freely choose an interaction approach (*push* or *pull*) that is most appropriate with respect to the market participant's internal processes (*R10*). Furthermore, the process of generating reputation values can be adjusted by each market participant individually by specifying the set of user ratings to be considered, the actual aggregation function to be applied, and the final representation of the aggregated value (*R8*).

Reputation information about OTF providers (*R1*) is provided by the reputation system in a straight-forward manner. Users rate their satisfaction regarding the transaction with an OTF provider. These ratings are accumulated and aggregated by the reputation system and can be accessed by other users. The process of generating reputation information about composed services (*R4*) is similar. Users rate their satisfaction regarding the execution process and the execution result of a composed service. These ratings, again, are accumulated and aggregated by the reputation system. In comparison to the reputation of OTF providers, however, reputation information about composed services is OTF provider-related. In order to preserve business secrets, only the OTF providers themselves can access the reputation values of their own composed services - after successful authentication (*R9*).

Besides being directly rated by users, ratings of elementary services also have to be derived from ratings for composed services (*R3*). For this purpose, OTF providers send information about their composed service to the reputation system. In order to not reveal their business secrets, this composition information, however, only consists of abstract, structural information. Only the set of elementary services included in a composed service is exposed, but not, for example, when and how often a particular service is called. This way, the provider's business secrets are protected, while it also allows for a mapping of the rating for a composed service to single services (*R6*).

Since users only interact with OTF providers, user ratings for service providers cannot be provided to the reputation system (*R6*). To overcome this problem, the aggregation system extrapolates from reputation information about elementary services to information about the associated service providers during the aggregation process (*R2*).

While composing services, reputation information about elementary services have most likely to be aggregated in order to choose composed services not only based on their (aggregated) non-functional properties, but also based on their overall reputation. How to determine this overall reputation, however, depends on the user requirements and the composition strategy of the respective OTF provider. If a user requires, e.g., all elementary services to satisfy a minimal reputation value, an OTF provider has to check the reputation value of each service individually. Another user might be satisfied with an average reputation value above a specific threshold. In this case, an OTF provider has to determine the average reputation value by aggregating all single values. Subsequently, the aggregated value and the threshold value have to be compared. In either case, aggregation of reputation values within the composition process is not part of the reputation system itself. A further investigation of how to integrate reputation information into service composition is addressed in Section VIII.

VI. Related Work

There is a lot of literature on reputation, both in economics and computer science. Our interpretation of reputation is used for instance by Shapiro [21] or as well by Bar-Isaac and Tadelis [22], who summarize the economic literature on reputation. Design aspects related to mathematically modeling a reputation system and challenges that arise with online transactions, are explicitly discussed by Friedman et al. [23] and Dellarocas [24], for example. Another comparison of trust and reputation models without referring to services or

service composition is given by Gómez Mármol and Martinez Pérez [25].

More closely related, we identify three involved fields, *Reputation Systems*, *Privacy-Preserving Systems* and *Service Composition*, and their overlaps with each other as shown in Figure 5. In the following, we present related work, which has been done within these overlaps in more detail. It is noteworthy that no work covers all of the three different fields (the overlapping marked **x** in Figure 5). To the best of our knowledge, we are the first to take all three fields into account.

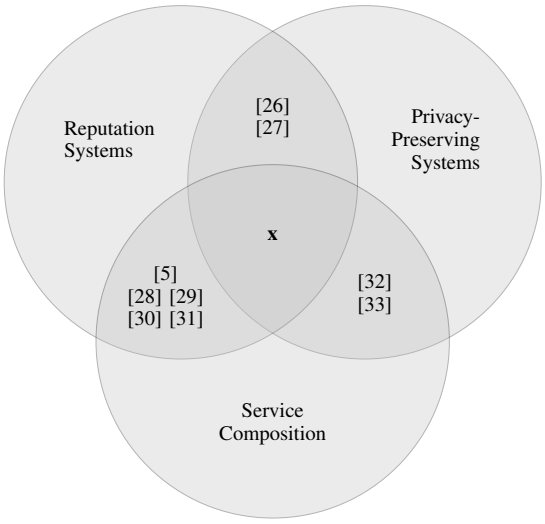


Figure 5. Overview of related work.

A. Reputation Systems and Privacy-Preserving Systems

Researchers have come up with *privacy-preserving* reputation systems in the past. Androulaki et al. [27] propose a reputation system that allows anonymous ratings. However, there is no authorization mechanism in their approach, i.e., anybody could rate any service. Kerschbaum et al. [26] present a system, which requires two centralized mutually mistrusting reputation providers in order to achieve anonymous user ratings. Users encrypt their ratings and send them to the first reputation provider, which collects a number of ratings and then publishes them to a bulletin board. The second reputation provider retrieves the ratings from the bulletin board to decrypt and aggregate them before providing a (computed) reputation value. The approach is based on the Paillier cryptosystem [20]. However, the approach is too inflexible and complex to be used in our OTF Computing setting. We want to keep a lean OTF infrastructure with only one reputation provider and no other additional components, such as a bulletin board, used only by the reputation system.

B. Reputation Systems and Service Composition

In general, existing approaches in this area focus on technical issues, e.g., how exactly reputation of composed services is computed. Our work does not focus on these issues but rather provides a holistic overview of how reputation is used in different processes of OTF Computing.

A good overview of existing research in the crosscutting area of reputation systems and service composition is provided by Gómez Mármol and Kuhnen [5]. In their survey, they compared 12 approaches on reputation-based web service composition. According to their analysis, most of the considered approaches suffer from security issues like sybil attack. Furthermore, many of those approaches offer some flexibility in a way that they support different aggregation strategies. The latter is similar to our approach. However, these approaches are not applicable to OTF Computing because they do not consider reputation of the different roles, i.e., OTF Provider and Service Provider. Also other requirements are not covered, e.g., context-specific reputation, expert ratings, or different ways of interaction. Moreover, the matching approaches used in the considered approaches to select services for a composition based on reputation are rather simple and do not support complex requests or different fuzziness sources.

Ali et al. present a reputation system with an integrated *Service Composer* [28]. They combine reputation with service composition by evaluating reputation metrics whenever services are composed. The *reputation computation phase* calculates reputation for elementary services as well as for composite services. In another approach, Motallebi et al. integrate *Component Reputation* and *Component Trust* in order to derive the reputation of a composed service from trust values for single services [29]. They do this by taking into account the frequency of invocations of these services. Both approaches covers only some of our requirements for a reputation system in OTF Computing. For example, neither service providers are considered, nor is privacy or security a topic within their publications.

Malik and Bouguettaya present the framework RATEWeb [30], which aims at facilitating service composition considering a service's reputation. In this approach, the service comsumer is responsible for maintaining reputation values, i.e., the reputation system is distributed. This contradicts with our idea of OTF providers that use "global" reputation values within service composition. However, their reputation metrics are very comprehensive in a way that they consider different aspects of computation, e.g., rater credibility, majority rating, and temporal sensitivity. We focus on a more flexible method to be configured by the user at runtime. In future, we should cover a similar range of aspects in our requirements, though.

The reputation propagation framework by Huang et al. [31] considers "various entities in the [service] ecosystem", e.g., not only a service's reputation but also the providers' reputation is considered. This makes it similar to our approach. However, they do not describe how reputation is considered within other processes, e.g., service composition. Furthermore, the matching process itself is rather simple, based on one numerical value. An interesting idea is that they take the domain of a service into account. This might also be an interesting addition to our approach in the future.

C. Service Composition and Privacy-Preserving Systems

Tbahriti et al. identify privacy preservation as one of the most challenging problems in *Data-as-a-Service (DaaS)* services composition [32]. DaaS is about combining web services for data publishing and sharing. In their proposed approach, *privacy policies* specify how collected data is treated and *privacy requirements* specify how the service-consuming

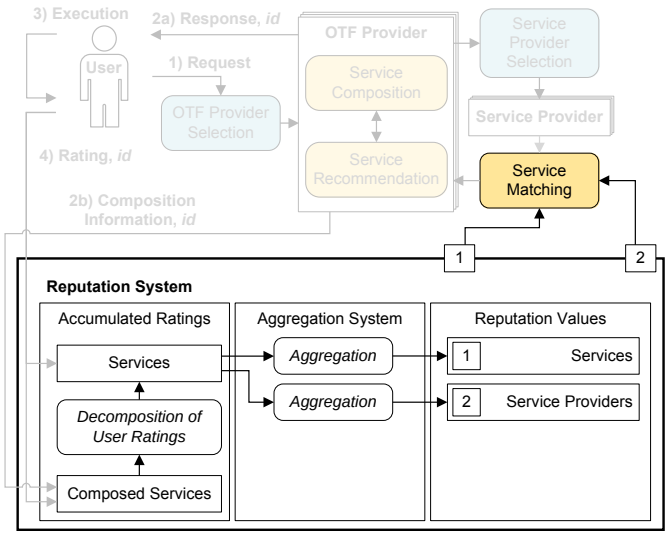


Figure 6. Overview: Service Matching

services are expected to treat the provided data. Similarly, Costante et al. come up with a solution for web service selection and composition that takes privacy into account [33]. Users are able to specify their privacy preferences, which are checked against the service providers' privacy policies. Only in the case of a successful match are the service providers' services selected and used for composition. Both approaches do not take into account reputation of elementary or composed services.

In contrast to related work, we pursue a *privacy-by-design* approach that builds privacy protection into the reputation system for OTF Computing. This allows us to prove that privacy is achieved rather than to rely on guarantees made by the participants.

VII. Fuzzy Reputation Matching

In general, service matching is the process of comparing a request for a certain service to descriptions of the services provided in a service market. For each provided service, a matching process delivers a matching result that indicates how much the service satisfies the given request. Here, we do not distinguish between composed services and elementary services: the provided service can be of both types. The matching process includes functional properties, e.g., signatures or protocols, as well as non-functional properties, e.g., quality properties of the service, like performance. A service's reputation is part of the non-functional properties. The most important difference between reputation matching and matching of other service properties is the source of the information the request has to be matched to: While properties like signatures and protocols have to be specified by the service providers themselves, reputation data has to be determined by a reputation system managed by a trusted third party – the OTF provider in our case. This is due to the fact that the ratings a reputation value is aggregated from are security- and privacy-sensitive (see Section IV). The provider must be prevented from manipulating such data (also covered by *R12*).

Section V-B illustrated a very simple reputation matching example. In practice, we expect requesters to require a more complex matching approach. The reason is that their requirements, on the one hand, can be more much complex but, on the other hand, they can also be fuzzy. Similar facts hold for the information provided in the reputation system. In the following, we explain how a more complex reputation matching works, which kinds of fuzziness can occur, and how we cope with them. As depicted in Figure 6, in this section, we only focus on the matching process and not on the process that gathers the ratings needed to calculate reputation. The reputation matching process depends on reputation values for services and service providers.

A. Reputation Matching

A request for a service's reputation consists of a list of conditions that can be fulfilled or not. As an example, consider the request in Figure 7. This request consists of four conditions, `c1` – `c4`. These conditions can be evaluated based on data from the reputation system. For a full match, all four conditions have to be true. If not all conditions are true, the matching approach returns a result that denotes to what extent the request is satisfied. Based on this result, the requester (or the OTF Provider's recommendation system, see Section VIII) can compare and select between different services. The more complex a request is, i.e., the more details a requester specifies, the more accurately can the matcher determine results that actually fit the requester's interests. However, with an increasing complexity, also the set of services matching the request to a high extent becomes smaller. We explain the example request depicted in Figure 7 in more detail in the following.

Each of the conditions in a request checks several properties related to service reputation (*R3*) and a service provider's reputation (*R2*). For example, `c1` checks whether the overall reputation of a service is greater or equal to `4` (based on a five star range as it is common in today's app stores). As a further restriction to this reputation value, this value must have been aggregated based on at least 100 ratings. Such restrictions are useful as a reputation value's reliability increases with the number of ratings it has been calculated from. The conditions `c2` and `c3` check context-specific reputation values (*R5*), i.e., the reputation with respect to the perceived response time of the service (`c2`) and the reputation with respect to the perceived security of a service (`c3`). Furthermore, `c2` uses an approximation operator ($\approx$). It means that the lower bound should be approximately `4` but the requester is tolerant regarding slight deviations. For example, a reputation of `3.95` would also be accepted. We will give more details about such approximated conditions in the next subsections related to fuzzy matching. In `c3`, we can see a restriction with respect to time. In this example, the reputation value should have been created based on at least 50 ratings, which have been given during the last three months. These kinds of restrictions are based on the idea that recent ratings are more reliable than old ones. This especially happens if the rated service has been updated or if the environment of the raters changed (e.g., the global sensitivity to security increased due to some incident). In contrast to `c1` – `c3`, `c4` is about the reputation of the service's provider. Another special characteristic of `c4` is the restriction `newer > older`. Based on the same idea

Request:

c1: Rep(Service) ≥ 4
based on ≥ 100 ratings

c2: Rep_{RT}(Service) ≈ 4
based on ≥ 50 ratings

c3: Rep_{Sec}(Service) ≥ 3
based on ≥ 50 ratings
of the last 3 months

c4: Rep(Provider) ≥ 3
newer > older

Reputation System:

Entity	Context	# Ratings	
Service: ImagePro1	Overall	300	Rep(300 ratings) = 4.5
Service: ImagePro1	ResponseTime	80	Rep (80 ratings) = 3.5
Service: ImagePro1	Security	30	Rep (30 ratings) = 2
Provider: IPServices	Overall	1300	Rep (newer > older) = 4
...	...	...	...
Service: PicProcessor	Overall	90	Rep (90 ratings) = 3.35
Service: PicProcessor	ResponseTime	30	Rep (30 ratings) = 1.5
Service: PicProcessor	Security	30	Rep (30 ratings) = 4.5

Figure 7. Exemplary request and extract of reputation system contents

as explained above, this restriction is related to the age of the ratings the creation of the reputation value is based on. However, in this case, the requester did not specify a specific threshold as above. Instead, ratings are weighted based on their age. For example, ratings from the current month have a higher impact on the reputation value than ratings given a year ago. Please note that flexible feedback processing (*R8*) is an essential prerequisite for such complex requests as they affect the aggregation function the evaluated reputation values are based on.

The right part of Figure 7 shows an exemplary extract of the contents of a reputation system in a tabular notation. These contents, amongst others, are used to evaluate the conditions of the request explained above. For example, reputation values based on different contexts for the services `ImagePro1` and `PicProcessor` are depicted. We also see the reputation of the service provider `IPServices`, which is the provider of both the `ImagePro1` and the `PicProcessor` services. The third column depicts how many ratings are available per service in total. The rightmost column depicts some exemplary reputation values calculated based on these ratings. Please note that these are dynamic values not stored in the reputation system but derived from the ratings stored in the system based on a selected aggregation function.

After the required reputation values considering all requested restrictions have been determined, an exact matching of each condition is a simple numerical comparison. For example, for `ImagePro1`, `c1` evaluates to `true` because the overall reputation value can be calculated based on 300 ratings and turns out to be 4.5 (cf. the first row of the reputation system depicted in Figure 7). In contrast, `PicProcessor` can already be sorted out as the overall reputation can only be determined on 90 ratings (and it is only 3.35, anyway). From this example, we can learn that an exact matching approach approach like this always depends on complete knowledge, e.g., the requested number of ratings. In the following, we explain why this is an unrealistic assumption and how we can deal with incomplete knowledge using fuzzy matching.

B. Fuzziness Types

Since the reputation of a service is not an objective measure, such as signatures or protocols, uncertainty or *fuzziness* might easily be introduced into the matching process. The more detailed the request, the more possibility there is for induced fuzziness.

Fuzziness can be introduced into the matching process due to several reasons. In our earlier work, we classified these reasons into different fuzziness types [11][12] including Requester-induced Fuzziness and Provider-induced Fuzziness:

- *Requester-induced Fuzziness:* Requesters are often tolerant with regard to slight variations between the stated requirements and the provided service. Especially "soft" constraints, like the conditions in a reputation request, are likely to be vague. For example, a requester wanting a reputation of 4, might also be satisfied with a service having a reputation of 3.95 if all other properties match well.
- *Provider-induced Fuzziness:* We expect service providers to not provide all information a service matching approach needs to determine an exact match in most cases. Reasons for this include (a) they do not want to publish many details in order to protect business interests and (b) they cannot know all required service properties because these details are difficult to determine, e.g., the overall performance of a service. Regarding reputation matching, provider-induced fuzziness occurs if the reputation system does not have the data needed to evaluate a condition on a given request. For example, if a service is rather new in the market, there cannot be many ratings for this service. Another reason might be unrateability of the provider or of a service that has only been used as part of a composition (*R6*).

Our main idea is to enable matching despite these uncertainties and to make them visible to the requester. Thus, the matching approach should not only return a matching result but also a measure of each of the three fuzziness types. As a benefit, the requester can make a more informed decision. In most cases, requesters will prefer services with good matching results that come with a low amount of fuzziness. Furthermore, if the requester-induced fuzziness is high, the requester can even react to it by modifying his request and restarting his search. Similarly, even the provider can try to react if he finds that matching results returned for his services are often inflicted with a high amount of provider-induced fuzziness in order to increase its sales opportunities.

There are different ways of how to represent a matching result that reflects induced fuzziness. For example, in our previous work, we assigned percentage fuzziness scores to privacy policy matching [34]. In the following, we will show an

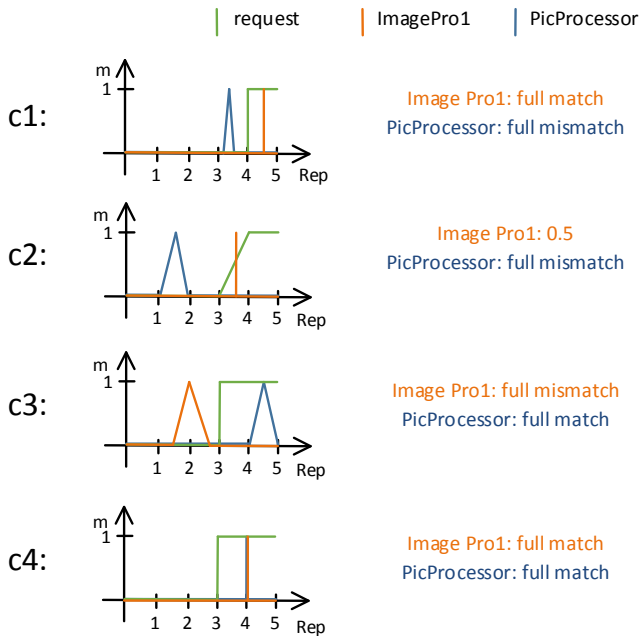


Figure 8. Exemplary Matching Results

approach appropriate for reputation matching based on fuzzy logic, more specifically fuzzy sets [35], as fuzzy logic has already been shown to be useful in order to model uncertainty in many works (e.g., [36], [37]).

C. Calculation of Fuzzy Matching Results

The example request depicted in Figure 7 covers requester-induced fuzziness as well as provider-induced fuzziness. Requester-induced fuzziness occurs in `c2`, indicated by the approximation operator $\approx$. When analyzing the service `ImagePro1`, provider-induced fuzziness occurs in `c3` because the reputation system does not have 50 ratings regarding security of this service but only 30. When analyzing the service `PicProcessor`, provider-induced fuzziness occurs in `c1`, `c2`, and `c3` as there is only a little amount of ratings for this service.

In order to evaluate `c1` – `c4`, these conditions are transformed into membership functions. Conditions inflicted with fuzziness are fuzzified into fuzzy sets. Figure 8 depicts these sets. The x axes denote reputation values in a scale from 0 to 5, while the y axes represent the membership as a number between 0 and 1. The sets for the request are depicted in green color. For example, the threshold for the requested reputation in `c1` is 4. Thus, the membership is 0 from 0 to 4 and 1 between 4 and 5. This means, if a service's reputation value is higher than 4, it matches completely. For `c3` and `c4`, the thresholds are 3. `c2` is transformed into a fuzzy set as there is no hard threshold. Depending on the risk affinity of the requester, the transition can be more or less steep. However, for simplicity reasons, at the moment, we model the steepness based on a constant value of 1. Thus, in this example, we modeled the fuzzy constraint with a smooth transition between full and no membership from 3 to 4. Modeled as a tuple, a fuzzy set can be denoted by its lower left corner, its upper left corner, its upper right corner, and its lower right corner. Accordingly, this fuzzy set is denoted by the values (3,4,5,5).

The orange and the blue lines are the membership functions for the two provided services to be matched (`ImagePro1` and `PicProcessor`). For example, the reputation value of `ImagePro1` is 4.5. This leads for `c1` to a membership of 1 at 4.5 and 0 otherwise. As stated above, for `PicProcessor`, there are not enough ratings to evaluate `c1`, so we model the provided reputation value as a fuzzy set. The fuzzy set is derived from the information we have and from the amount of missing information. In this example, we know that the reputation is 3.35 based on 90 ratings. It is uncertain how the reputation value would develop with 10 more ratings (a value based on 100 ratings was requested). However, assuming an averaging aggregation function for the reputation, the value cannot deviate much. Thus, based on the ratio of missing ratings and available ratings, we span a triangular fuzzy set. The peak is the value we calculated based on the currently available ratings and the transition, i.e., the steepness on both sides is determined by the number of missing ratings and the number of requested ratings. In this example, the peak is at 3.35 and the steepness on both sides is 0.1 (10 missing ratings out of 100 ratings requested: $10/100$). This leads to a fuzzy set of (3.25,3.35,3.45). In `c2`, there is more uncertainty about the reputation value of `PicProcessor` as even more ratings are missing compared to the request: The reputation value calculated based on the currently available ratings is 1.5, 20 ratings are missing and 50 are requested. In this case, we use the current reputation value of 1.5 and a deviation of 0.4 ($20/50$), which denotes a fuzzy set of (1.1,1.5,1.9).

Based on these membership functions, the provided services are compared to the request in order to decide whether the services match. Regarding `c1`, we have a full match for `ImagePro1` because its membership is fully covered by the request's membership function. `PicProcessor` does not match with respect to `c1` because there is no overlap with the set represented by the request's membership function. The evaluation of `c2` shows that the reputation value for `ImagePro1` intersects the request's membership function at a value of 0.5 on the y axis. Accordingly, the matching result for this condition is 0.5. `PicProcessor` does not match `c2`. Regarding `c3`, `ImagePro1` matches, while `PicProcessor` does not match. Furthermore, both services match with respect to `c4`.

The results for the four conditions are aggregated to one final matching result per service. Taking the average again, `ImagePro1` matches with a result of

$$(1 + 0.5 + 0 + 1)/4 = 0.625\ ,$$

and `PicProcessor` matches with a result of

$$(0 + 0 + 1 + 1)/4 = 0.5\ .$$

As a conclusion, the requester should choose `ImagePro1`.

D. Configurability

In Section VII-C, we described how we derived membership functions from the request that describes a user's requirements on reputation. There are some configuration possibilities to adapt these functions. For example, the steepness of the

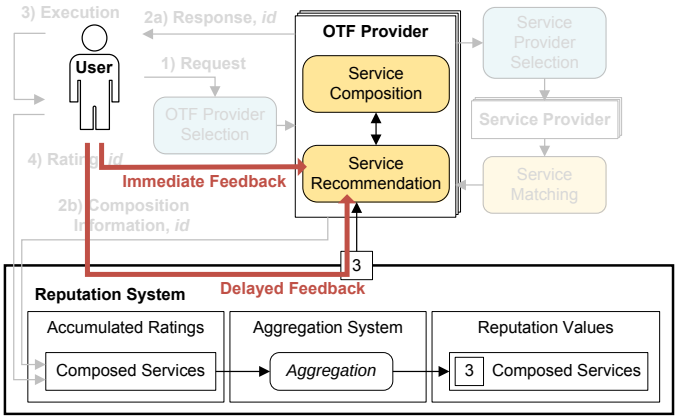


Figure 9. Overview: Service Composition and Recommendation

fuzzy sets can be determined based on different heuristics. Here, we only described one possibility. In general, choosing an appropriate heuristic means dealing with the trade-off between precision and recall: the steeper the transitions from membership to non-membership and vise versa, the less false negatives can be expected, while the less steep the transitions, the less false positives can be expected.

Another possibility for configuration is the aggregation function. Here, we described an averaging aggregation method. However, also maximizing or minimizing methods or more complex, e.g., hierarchical methods [38], are possible.

VIII. Reputation for Composed Services

This section focuses on the internal processes of a single OTF provider while taking user feedback for composed services into account (see Figure 9). In this context, the service composition component realizes a sequential service composition approach (cf. Section VIII-A). The service recommendation component, in turn, realizes a learning approach in order to improve the quality of composed services over time (cf. Section VIII-B). In this section, quality of composed services corresponds to how good a composed service satisfies a desired functionality. Assuming that alternative realizations for a desired functionality exist, the alternative that approximates the desired functionality the best is interpreted as the composed service with the highest quality.

After briefly introducing the fundamental techniques of both service composition approach and service recommendation approach, we present experimental results for demonstrating how delayed feedback in terms of reputation values influence the learning behavior, which originally incorporates immediate feedback. We use the image processing example introduced in the beginning of this paper as concrete use case for composing, executing and rating composed services. Incorporating reputation values for single services, however, is not considered in this section.

A. Sequential Service Composition Model

As already stated in Section III-A, we interpret automatic service composition as the sequential application of composition steps. In this section, we introduce our sequential composition model based on composition rules. However, our composition model represents only one possible realization of the service composition component depicted in Figure 9.

A composition rule compactly defines a formally correct modification during the service composition process. The syntax of composition rules is identical to the syntax of production rules, which contain *non-terminal* symbols and *terminal* symbols [39]. In our context, non-terminal symbols correspond to functional parts of the composed service that still have to be realized. Terminal symbols correspond to concrete services. For example, the composition rules $X \rightarrow s_1 Y \mid s_2$ define that a required functionality X can be composed by a service s_1 and a required functionality Y. Non-terminal Y, in turn, has to be realized in a subsequent composition step. Alternatively, X can be directly realized in terms of service s_2.

For illustration purposes, let us consider the image processing example introduced in Section II. Let us assume that Figure 1a can be transformed into Figure 1b by a sequence of two post-processing services. In fact, four different steps are actually necessary to achieve the desired result: Modification of contrast, brightness, saturation, and sharpness. However, for reasons of comprehensibility and without loss of generality, we consider only two processing steps Y and Z to be necessary in order to achieve the desired functionality X. The chronological order, in which both steps are applied cannot be neglected, but is relevant due to dependencies between processing results. In short, YZ may produce different results then ZY. As a consequence, we write $X \rightarrow YZ \mid ZY$. Let us assume, that each processing step can be performed by two different services, resulting in four services: s_1, s_2, s_3, and s_4. Services s_1 and s_2 as well as services s_3 and s_4 implement similar functionality and are equivalent with respect to their formal specifications. Required functionality Y can be realized by either s_1 or s_2, while Z can be realized by either s_3 or s_4. In short, we write $Y \rightarrow s_1 \mid s_2$ and $Z \rightarrow s_3 \mid s_4$.

Figure 10a depicts the complete list of composition rules mentioned heretofore. Each rule is assigned a unique identifier $\widehat{r}_i$. A derivation of rules corresponds to composing a solution for desired functionality X. For example, the derivation

$$X \xrightarrow{\widehat{r}_2} ZY \xrightarrow{\widehat{r}_3} Zs_1 \xrightarrow{\widehat{r}_5} s_3 s_1$$

composes the solution $s_3 s_1$ by successively applying rules $\widehat{r}_2$, $\widehat{r}_3$, and $\widehat{r}_5$. In fact, this sequence of single decision-making steps is similar to workflow composition. First of all, an abstract workflow has to be selected by either applying rule $\widehat{r}_1$ or rule $\widehat{r}_2$. Afterwards, the functional placeholders Y and Z have to be realized by choosing between rules $\widehat{r}_3$, $\widehat{r}_4$, and rules $\widehat{r}_5$, $\widehat{r}_6$, respectively.

According to Rao and Su [40], the second major realm of approaches besides workflow management that have emerged in order to tackle the service composition problem is Artificial Intelligence (AI) planning. From the AI planning perspective, service composition can be interpreted as a search problem in a state transition system [41]. Figure 11a depicts the state space for our example. In the most general sense, a state s_i corresponds to a set of conditions that hold as long as state s_i is occupied. Actions (arrows), in turn, correspond to services. Applying a service to a state s_i corresponds to changing the

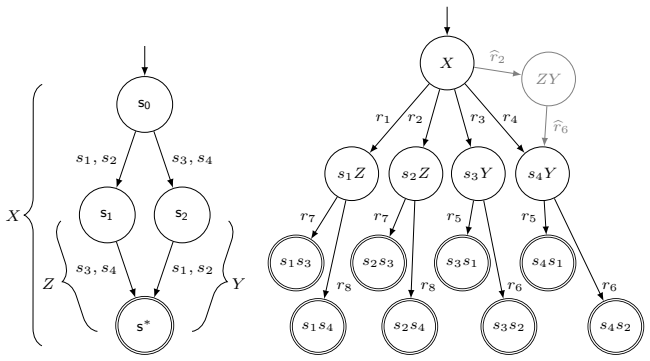
	$\widehat{r}_1$	$\widehat{r}_2$
X	YZ	ZY
	$\widehat{r}_3$	$\widehat{r}_4$
Y	s_1	s_2
	$\widehat{r}_5$	$\widehat{r}_6$
Z	s_3	s_4

(a)

	r_1	r_2	r_3	r_4
X	s_1Z	s_2Z	s_3Y	s_4Y
	r_5	r_6		
Y	s_1	s_2		
	r_7	r_8		
Z	s_3	s_4		

(b)

Figure 10. Composition rules (a) according to workflow composition and (b) according to forward search from AI planning.



(a) Planning State Space (b) Composition Rule State Space

Figure 11. Relationship between planning based composition and sequential application of composition rules.

conditions encoded in state $\mathsf{s_i}$ [42]. A desired functionality is usually specified in terms of pre- and postconditions. The preconditions correspond to the initial state $\mathsf{s_0}$, while the post-conditions correspond to the goal state $\mathsf{s_*}$. Search algorithms are then applied to find a sequence of services that transforms $\mathsf{s_0}$ into $\mathsf{s_*}$.

The planning based composition model can be gradually transformed into our composition model during the search process [43]. In terms of forward search, a non-terminal symbol N represents the unresolved path from the currently occupied state $\mathsf{s_i}$ to the goal state $\mathsf{s_*}$; we write $N = \overrightarrow{\mathsf{s_i s_*}}$. For example, when the forward search enters state $\mathsf{s_0}$, non terminal symbol $X = \overrightarrow{\mathsf{s_0 s_*}}$ is constructed. If the forward search then proceeds to state $\mathsf{s_2}$ by applying service s_3, non-terminal symbol $Y = \overrightarrow{\mathsf{s_2 s_*}}$ and composition rule $r_3 : X \rightarrow s_3 Y$ are constructed. If the search process subsequently applies service s_1 and proceeds to state $\mathsf{s_*}$, rule $r_5 : Y \rightarrow s_1$ is constructed. Figure 11b shows the complete state space based on our composition model for our example. The associated composition rules are listed in Figure 10b.

There are two major benefits when using our composition model. First, our model facilitates a unified composition process. Both workflow composition and AI planning can be modeled in terms of composition rules and can be combined in the same state space. Rule $\widehat{r}_3$, e.g., is equivalent to rule r_5. The sequential application of rules $\widehat{r}_2$ and $\widehat{r}_6$, in turn, is combined in rule r_4 (cf. Figure 11b). Second, the underlying decision-making process in our model satisfies the Markov property: Decisions do not depend on history but are memoryless. All relevant information is encoded in a single state in terms of the current composition structure. Whenever a decision between alternative composition rules has to be made, decision-making bases on the heretofore composed solution. Choosing between alternatives is up to the applied RL approach.

B. Improving Quality by Means of Learning

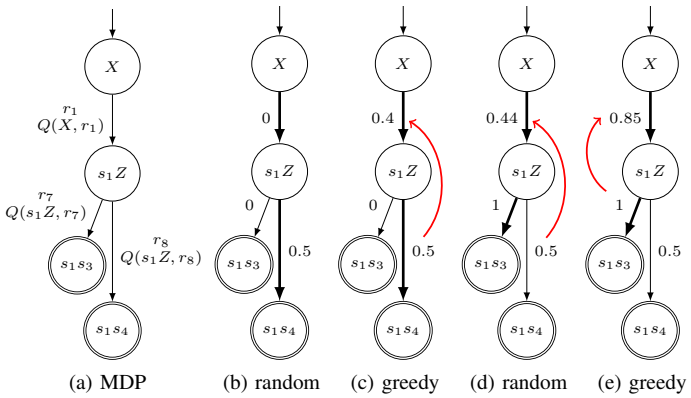
The recommendation component (cf. Figure 9) incorporates RL techniques for improving the quality of composed services over time. In general, RL addresses the problem faced by an autonomous agent that must learn to reach a goal through sequential trial-and-error interactions. Depending on its actions, an agent receives a reward value that is incorporated into the decision-making processes in order to adjust the selection of actions in the future. In case of episodic RL, an agent is not learning continuously but periodically in terms of episodes. An episode defines the period between initial state and final state.

In our context, an episode covers an entire composition process. After each composition process, a user gives feedback by rating the quality of the *execution result*. In terms of our example, a user rates his satisfaction regarding the result image that is produced by the composed service. To roughly simulate user ratings in our experiments, the original image (Figure 1a) is automatically processed by the composed service. The resulting image is then compared with the manually produced desired image (Figure 1b). Based on the normalized distance value $d \in [0, 1]$ of the original image and the desired image, the user rating $r = 1 - d$ is computed. The higher the rating, the better the execution result and, consequently, the higher the quality of the composed solution. The user rating is subsequently incorporated as a final reward into the sequential composition process to improve the selection of alternative composition steps during future composition processes.

Technically, the sequential composition process as described in the previous section is modeled as Markov Decision Process (MDP) [44]. Figure 12a shows a snippet of the state space depicted in Figure 11b. The annotated quality values $Q(\mathsf{s}, r)$ are an estimation of how good it is to apply a composition rule r when currently occupying state s. The higher these so-called Q-values, the more promising the corresponding actions for the composition task. In our work, we apply Q-Learning to adjust the Q-values over time based on user feedback [45]. Q-Learning is a model-free RL algorithm that directly approximates Q-values by means of its update function

$$Q(\mathsf{s}_t, r_t) \leftarrow Q(\mathsf{s}_t, r_t) + \alpha \left[\gamma \max_r [Q(\mathsf{s}_{t+1}, r)] - Q(\mathsf{s}_t, r_t) \right], \quad (1)$$

with current state s_t, next state s_{t+1}, current composition rule r_t, next composition rule r_{t+1}, discount factor γ, and learning rate α. In addition to the update function, we incoporate an ϵ-greedy action selection strategy. Action selection strategies are crucial in order to balance between exploitation of already learned knowledge and exploring new and probably better states. In case of ϵ-greedy, composition rules are greedily selected with probability $1 - \epsilon$: the composition rule with the highest Q-value is selected (exploitation phase). With probability ϵ, however, a composition rule is randomly selected (exploration phase). For our experiments, we chose a commonly used setting; that is, $\gamma = 0.9$, $\alpha = 0.9$, and $\epsilon = 0.1$.



(a) MDP (b) random (c) greedy (d) random (e) greedy

Figure 12. Demonstration of Q-Learning.

Figures 12b-12e illustrate the actual learning process. Each figure shows state space and associated Q-values *after* a composition process was completed and a user rating was incorporated as final reward. Thick arrows indicate the chosen path from initial state to final state. Q-values $Q(X, r_1)$, $Q(s_1Z, r_7)$, and $Q(s_1Z, r_8)$ are initialized with 0.

Figure 12b: Service s_1s_4 was composed and executed. During the composition process, composition rule r_8 was chosen randomly. The execution result was rated with value 0.5. The rating value was immediately integrated as final reward by adjusting Q-value $Q(s_1Z, r_8)$. Note: Final reward is always incorporated unmodified and replaces the Q-value of the lastly applied composition rule.

Figure 12c: The composition process again produced composed service s_1s_4. Composition rule r_8, however, was not selected randomly, but greedily based on Q-value $Q(s_1Z, r_8)$ that was modified in the previous composition process. After selecting r_8 and before transitioning to state s_1s_4, update rule (1) is applied to adjust Q-value $Q(X, r_1)$. Q-value $Q(s_1Z, r_8)$, however, is not changed again, since it is equivalent to the rating result, which is the same as before.

Figure 12d: Composition rule r_7 was randomly selected during the composition process. Executing composed service s_1s_4 results in an image that is identical to the desired result image. Hence, the rating value is 1. Q-value $Q(s_1Z, r_7)$ is immediately updated. During the composition process, however, this value was not yet available. Due to the max operator, Q-value $Q(X, r_1)$ was again updated based on Q-value $Q(s_1Z, r_8)$.

Figure 12e: The composition process operated in a greedy manner again. Furthermore, Q-value $Q(X, r_1)$ significantly increased, since it was updated based on Q-value $Q(s_1Z, r_7)$ this time.

By consecutively applying the update rule when moving through the state space and by continually incorporating ratings of consecutive composition processes, ratings are propagated throughout the state space and Q-values finally converge. Generally speaking, the overall composition process adapts its composition strategy to produce a composed service that approximates the desired functionality that is implicitly determined by the feedback.

We conducted two initial experiments to obtain reference results before investigating the impact of reputation information in the subsequent experiments. The objective in each experiment was to identify the composed service that reproduces the desired image shown in Figure 1b at best. Four different image processing operations had to be applied in order to appropriately modify contrast, brightness, color, and sharpness, respectively. For each operation, six different services with slightly different functionality were provided. The chronological order of the four operations was not defined in advance. To obtain sound results, each experiment comprised 3000 consecutive composition processes and was repeated 100 times. For each experiment, the ratings per composition process were plotted in terms of mean value and 95% confidence interval of all 100 independent runs. For reasons of clarity, the resulting plots were additionally smoothed.

Figure 13 depicts the results of the two initial experiments. In the first experiment (green plot), no feedback was provided at all. Decisions between alternative composition steps were made only randomly. As a consequence, the rating values do not increase over time but remain almost the same. The results of the first experiment serve as worst case. In the second experiment (black plot), immediate feedback was incorporated as described above. Rating values were directly integrated into the composition process as soon as they were available. The impact of learning is clearly visible. The period, in which the rating values increased, indicate the phase, in which Q-values kept changing during the experiment – the actual learning period. After approximately 1200 composition processes, the Q-values converged. The optimal value 1 was not achieved due to the ϵ-greedy action selection strategy: Composition steps were still selected randomly once in a while. The results of the second experiment serve as best case for the Q-Learning setting in our example.

C. OTF Provider-related Reputation Values

Considering our proposed reputation system, user ratings are accumulated and aggregated into reputation values to ensure preservation of privacy (cf. Section V-C). Furthermore, we consider reputation information about composed services to be OTF provider-related in order to strictly enforce preservation of OTF providers' business secrets. To evaluate the impact of OTF provider-related reputation values on the learning process, we delayed the incorporation of our simulated user ratings.

Usually, when dealing with markets of composed services,

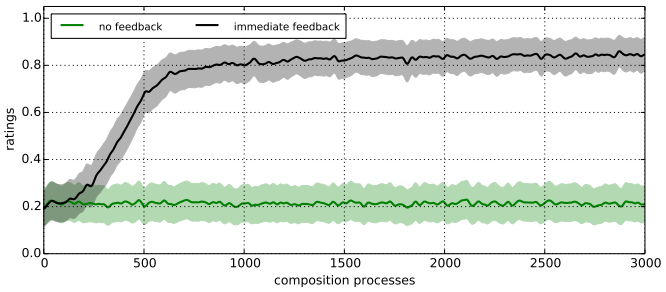


Figure 13. No learning (green) vs. learning (black).

users have individual preferences and consequently deliver different ratings for the same composed solution. By analyzing the different ratings, e.g., a new learning context can be identified and an independent learning process with a different learning objective can be initialized. To cope with the privacy preservation mechanisms, an aggregation function that still allows these steps has to be found. In our next experiment, however, we assumed that ratings are identical for identical composed services in order to analyze the mere influence of delayed reward. We conducted the experiment with the same settings as in the experiment with immediate feedback. However, we delayed the incorporation of ratings until a second rating for the same composed service was available. Figure 14 shows the results (blue plot) in comparison to the best and the worst case, respectively. The impact is enormous. After 3000 composition processes, the results of the best case were still not achieved. Nevertheless, a learning behavior is still clearly visible.

A possible way to improve the learning behavior is to change the learning process itself, including the update rule as well as the action selection strategy. For example, the update rule could be applied more often in order to increase the propagation speed of Q-values within the MDP. Modifying the learning process, however, is beyond the scope of this paper, but needs a more thorough investigation. Another common mean is to provide more samples, from which the process can learn. In our context, a sample comprises a composed service and its corresponding rating value. If more samples were available, reputation values could be provided earlier and the impact of the delay would be weakened. Regarding our OTF Computing market, OTF providers might establish a cooperation in order to share business secrets (the ratings for services they composed). The reputation system can be extended to accumulate and aggregate ratings for composed services that belong to a group of OTF providers. A new challenge that emerges is the asynchronous provision of new feedback. Until now, we assumed that feedback is incorporated synchronously – after a composition process has finished. However, new ratings from other OTF providers may be available anytime.

D. Publicly Accessible Reputation Values

In the previous section, we assumed that the complete composition information is stored in the reputation system in order to unambiguously identify and assign ratings to identical compositions. An alternative approach to increase the number of samples is to make rating values for composed services publicly accessible for every OTF provider. To not reveal all business secrets of the OTF providers, information about composed services that is stored in the reputation system must be abstracted. As a consequence, user ratings for different composed services most likely end up in the same accumulation process. When, for example, only information about the utilized services is considered, but not information such as data and control flow, ratings for composed services that contain the same set of services are mixed up. We call this effect ambiguous feedback.

To investigate the impact of ambiguous feedback, we conducted three additional experiments with different delays. During these experiments, composed services that contained the same set of services were considered to be identical. As a consequence, more ratings for a composed service were accumulated over time. The ratings for actually identical composed services, however, were not identical anymore. For example, the ratings for the composed service $s_1 s_2 s_3 s_4$ were accumulated together with the ratings for the composed service $s_4 s_3 s_2 s_1$. According to the defined delay, we aggregated the latest ratings in terms of their average. This value was then incorporated as ambiguous feedback into the learning process.

Figure 15 compares the results of the previous experiment and the first experiment with ambiguous feedback. In both experiments, reputation values were not provided until two rating values were available. At first view, the results were quite unexpected. The learning process with delayed, ambiguous feedback (orange plot) converged significantly faster than the learning process that incorporated just delayed feedback (blue plot). In fact, the experiment with ambiguous feedback even shows a very similar learning curve like the best case experiment with immediate feedback (black plot); shifted to the right by a several hundred composition processes. The reason for the outcome of this experiment is most likely the special characteristics of our chosen image processing experiment. The chronological order, in which the provided services were executed in order to achieve the chosen goal was not as important as assumed. As a consequence, the ratings for the set of services that were utilized in a composed service tended to be very similar. Although still delayed, feedback is accumulated faster and not as ambiguous as assumed.

Figure 16 compares the results of all three experiments with ambiguous feedback. The results of the experiment with

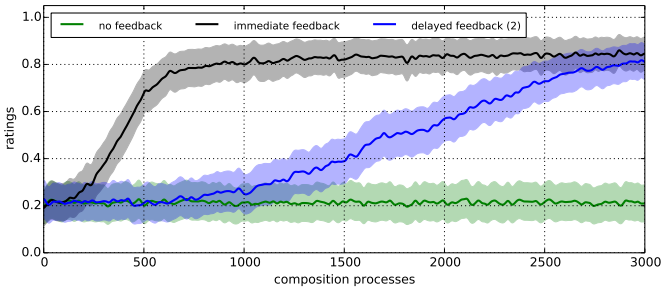


Figure 14. No learning (green), learning with immediate feedback (black), and learning with delayed feedback (blue) until two ratings were available.

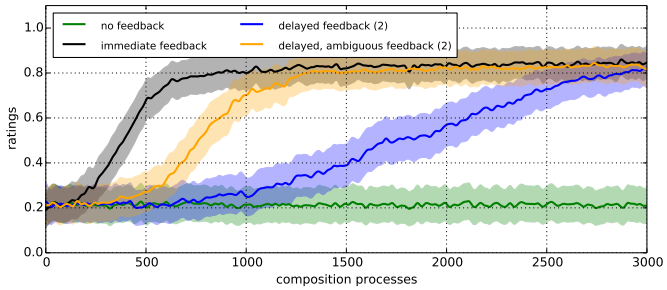


Figure 15. Delayed feedback (blue) based on two OTF provider related ratings and ambiguous feedback (orange) based on two publicly accessible ratings.

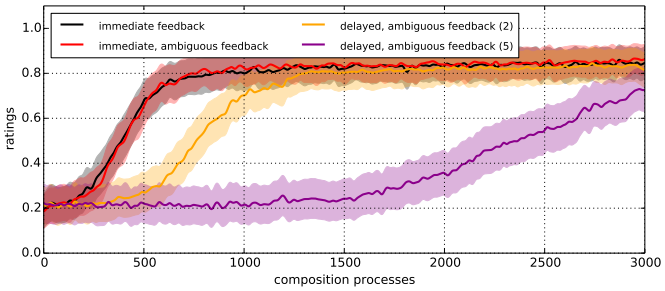


Figure 16. Comparison of all experiments with ambiguous feedback.

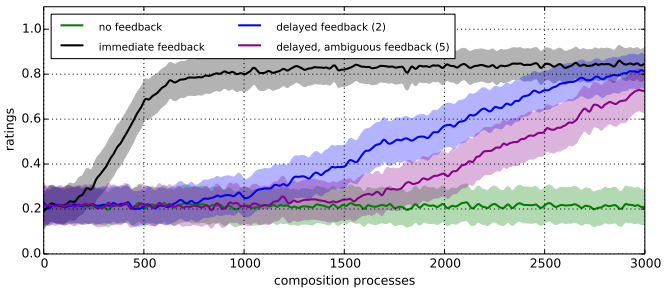


Figure 17. Comparison of delayed feedback (blue) based on two OTF provider-related rating values and ambiguous feedback (magenta) based on five publicly accessible rating values.

immediate, ambiguous feedback clearly show that mixing up ratings for composed services that are considered to be identical as long as they include the same set of services has no significant negative effect at all. The learning curves of both the experiment with immediate feedback and the experiment with immediate, ambiguous feedback do not significantly differ. Even the results of the last experiment (magenta plot), in which reputation values are not accessible until 5 rating values are available seem to be acceptable when directly comparing them to the delayed but OTF provider-related ratings (blue plot), as shown in Figure 17.

E. Remarks

In order to focus on the influence of reputation information under optimal conditions, we only considered a static context during the experiments. Neither the services available on the market, nor the specific user request changed. If we assume a MDP with a finite state space, Q-Learning is theoretically able to identify the optimal policy — as long as states are visited infinitely often [46]. The learning speed in terms of convergence rate can be improved by implementing a more sophisticated action selection strategy such as a dynamic ϵ-greedy strategy, where ϵ is adjusted based on temporal information.

In our case, however, we deal with a theoretically infinite state space: A composed service can be of arbitrarily length. In cases with low complexity such as our specific image processing example, the optimal solution may still be identified or at least be well approximated, as the experimental results have shown. In more complex scenarios, however, the state explosion problem inevitably emerges. A possible solution to deal with this problem is a dynamic state space approach. Similar states are merged into a single abstract state, while abstract states are split up if necessary. As a consequence, the amount of states can be restricted. Furthermore, in case of an abstract state, Q-values do not apply to a single concrete state, but to a set of states. This generalization effect might also be exploited to improve the learning speed [13].

In the long run, OTF Computing intends to consider dynamic market environments. Services may enter or leave a market anytime. In this case, finding the optimal solution is hardly possible. In fact, the Q-Learning approach will most likely only converge temporarily. However, finding the optimal solution (best case) is not even necessary to provide users a composed service that is better than a randomly selected solution. Users already benefit from less optimal solutions that are identified during the learning phase.

IX. Economic Decisions including Reputation

This section demonstrates how the information provided by the reputation system enters into the economic decision problems of users, OTF providers and services providers and how this influences the interaction in the OTF Computing market. Users and OTF providers interact when a user sends his request to an OTF provider (*Provider Selection*). OTF providers interact with service providers to purchase elementary services for a user's request (*Service Provider Selection*). This is shown in Figure 18. We suppose here that users and service providers do not directly communicate with each other and every interaction takes place via an OTF provider. Each of the market participants faces a different economic optimization problem. The primary goal of a user is to find the most preferred composed service(s) at an appropriate price whereas OTF and service providers are interested to maximize their profits by trading composed or single services.

An important feature of the OTF Computing market are information asymmetries between the market participants on qualitative characteristics of services and composed services.

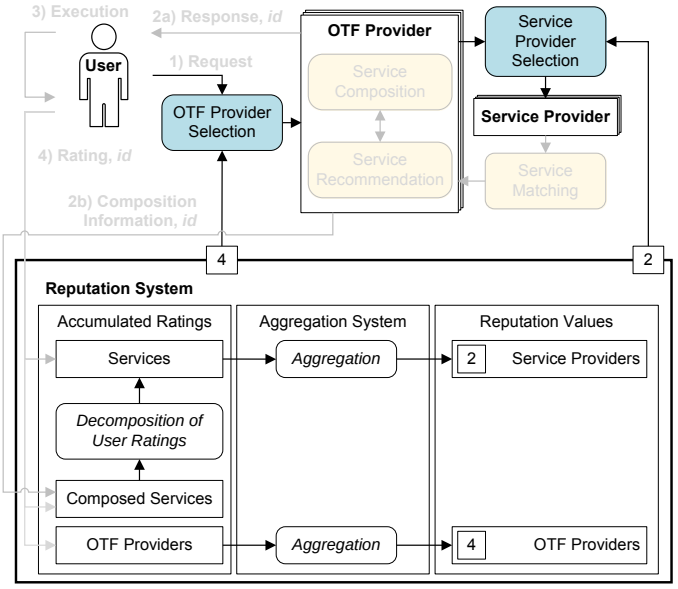


Figure 18. Overview: Economics

Therefore, a reputation system is essential to reduce these information asymmetries and to keep track of past behavior related to these unobservable characteristics. The information provided by the reputation system influences strategic decisions of users, OTF providers, and service providers. More precisely, the reputation information has an impact on the users' demands for composed services and, thus, also on the OTF providers' demands for elementary services. We model market prices depending on reputation values by analyzing both the market participants' decision problems and their interaction in the OTF Computing market.

For the rest of this section, suppose there are $N = 1, ..., n$ OTF providers typically indexed by i and $M = 1, ..., m$ service providers typically indexed by j. In addition, we assume for the notational simplicity that each service provider provides exactly one elementary service, which is available at different quality levels.

A. Economic Decisions of Users

The economic decision problem of a *user* is to formulate his request such that he gets the best composed service he can afford. The user's preferences are defined on a decision set of available composed services $S = \cup_{i \in N} \left\{ S_1^i, S_2^i, S_3^i, ... \right\}$. Hereby, the "available services" may be interpreted user-specific being those services that a user is aware of and considers in his decision problem. This need not necessarily be all composed services in the market. The user's preferences may also be expressed by weighting different characteristics of composed services. In this case, the distance between two composed services is used to define the user's preferences. If every two composed services from the decision set S can be compared with each other (*completeness*) and if S_1^i is preferred over $S_2^{i'}$ and $S_2^{i'}$ is preferred over $S_3^{i''}$ implies that S_1^i is preferred over $S_3^{i''}$ (*transitivity*) for $i, i', i'' \in N$, then the user's preferences are called rational [47, p. 42]. From a user's point of view there may exist exactly one ideal composed service (single peaked preferences) or his preferences may be monotonically increasing or decreasing for certain characteristics. Often utility functions describing the user's preferences are used to formally analyze economic decisions. In the OTF Computing market the user's utility function

$$u : S \to \mathbb{R} \tag{2}$$

assigns to a composed services $S_\ell^i \in S$ a valuation $u\left(S_\ell^i\right)$.

Within the OTF Computing process, the information provided by the reputation system about OTF providers influences the users' purchase decisions (*Provider Selection* in Figure 18). A user only selects those OTF providers that are reliable to properly answer his request. Therefore, taking the reputation information into account influences the user's evaluation of a composed service and may have the effect that certain alternatives are no longer acceptable as the reputation values are not sufficiently high. Let R_{OTF} and R_{SP} be the sets of possible reputation values of OTF and service providers. The user's utility function including the OTF provider's reputation value is given by

$$u : S \times R_{\mathrm{OTF}} \to \mathbb{R}. \tag{3}$$

Moreover, market prices in the OTF Computing market can be defined depending on the current reputation values. The user's decision problem is to buy a composed service that maximizes his utility for a fixed monetary budget that he is willing or able to spent. Therefore, given the OTF providers' prices $\left(p_{S_\ell^i}\left(r_{\mathrm{OTF}_i}\right)\right)_{S_\ell^i \in S}$, that depend on his current reputation value $r_{\mathrm{OTF}_i} \in R_{\mathrm{OTF}}$, a user selects those composed services that maximize the valuation minus the price for a composed service over the set of available composed services $S_\ell^i \in S$, formally the user's objective is

$$\max_{S_\ell^i \in S} \quad u\left(S_\ell^i, r_{\mathrm{OTF}_i}\right) - p_{S_\ell^i}\left(r_{\mathrm{OTF}_i}\right). \tag{4}$$

If a user is willing to buy several units of a composed service, his utility function and the price may also depend on the quantity he buys.

Besides the price of a composed service, quality considerations play an important role when a user compares different composed services in the OTF Computing market. If composed services are available at multiple quality levels, the user's decision problem can be explicitly extended by a quality dimension. Let Q be the set of possible quality levels and let Q_i denote the promised quality of composed service S_ℓ^i. Then, the user's utility function is

$$u : S \times R_{\mathrm{OTF}} \times Q \to \mathbb{R}. \tag{5}$$

The user's valuation of service S_ℓ^i in promised quality Q_i is $u\left(S_\ell^i, r_{\mathrm{OTF}_i}, Q_i\right)$ and therefore a user takes into account that each composed service may be available in different quality levels. A seminal contribution for the interplay of price and quality considerations in case of monopoly is [48], for instance. In the OTF Computing market, this quality is assumed to be unobservable before the transaction actually takes place. Therefore, the reputation information may help the user to find an appropriate OTF provider.

Beyond incorporating quality and reputation considerations, a different challenge for a user may be not to reveal his entire preferences in order not to be exploited in the market. Therefore, when he interacts with an OTF provider, he has to decide how much information he is willing to communicate about his preferences. This is important when his request is posed as well as when feedback on the composed service is provided. The revelation of preferences and the according incentives from an economic perspective are closely related to technical issues of privacy protection (cf. challenge *Manipulation Resistance versus Privacy Protection* in Section X).

B. Economic Decisions of OTF Providers

An *OTF provider* establishes a link between service providers and users. Thus, an OTF provider has to simultaneously consider the users' demands in the market for composed services as well as the available elementary services in the input market. Typically, there is a huge number of users with heterogeneous preferences in the OTF Computing market. Thus, from an OTF provider's perspective, the total users' demand is an aggregation of the individual demands from all users in the market resulting from their optimal buying decisions.

The profit maximization problem of an OTF provider is two-fold: On the one hand, it is crucial for an OTF provider to understand the users' preferences, to deduce their willingness to pay and to sell his composed services at profit maximizing prices. On the other hand, an OTF provider has to consider the supply of available services including their prices, that may either be given in the market or need to be negotiated with the service providers. The OTF provider's challenge is to find a combination of elementary services that minimizes the OTF provider's input costs for composed services (*Service Provider Selection* in Figure 18) and that at the same time sufficiently satisfies the users' requests (*Provider Selection* in Figure 18). It may be profitable for an OTF provider not to offer the entire range of composed services, but to focus on particular groups of users. The payments finally received for composed services need to be such that the prices paid for the elementary services as well as the effort put into the service composition process can be compensated. Hereby, the prices are typically dependent on the current reputation values provided by the reputation system.

The decision problem of an OTF provider $i \in N$ can be formalized as follows. Assume that the current reputation values are given by $r_{\mathrm{OTF}_i} \in R_{\mathrm{OTF}}$ and $r_{\mathrm{SP}_j} \in R_{\mathrm{SP}}$ for all $j \in M$ and denote the vector of the service providers' reputation values by $r_{\mathrm{SP}} = \left(r_{\mathrm{SP}_j}\right)_{j \in M}$. Let S^i be the set of composed services OTF provider i is able to build. Suppose, the OTF provider uses the composition strategy k_ℓ^i chosen from the set of possible composition strategies K^i to produce a composed service $S_\ell^i \in S^i$. This requires the use of elementary services s_{ij} for $j \in M_{k_\ell^i} \subseteq M$. Technical details of the *Service Composition* and *Service Recommendation Process* can be found in Section VIII. Given the sales price $p_{S_\ell^i} : R_{\mathrm{OTF}} \to \mathbb{R}_+$ with $r_{\mathrm{OTF}_i} \mapsto p_{S_\ell^i}\left(r_{\mathrm{OTF}_i}\right)$, prices of the elementary services $p_{s_{ij}} : R_{\mathrm{SP}} \to \mathbb{R}_+$ with $r_{\mathrm{SP}_j} \mapsto p_{s_{ij}}\left(r_{\mathrm{SP}_j}\right)$ and costs of the service composition $c : S_{\mathrm{OTF}_i} \times K \to \mathbb{R}_+$ with $\left(S_\ell^i, k_\ell^i\right) \mapsto c\left(S_\ell^i, k_\ell^i\right)$ the decision problem of an OTF provider for a user's request is to choose a composed service $S_\ell^i \in S$ and a composition strategy $k_\ell^i \in K^i$ to maximize his profit. This is the price the OTF provider receives from the user minus the costs he has. Formally, OTF provider i's profit is

$$\pi_{\mathrm{OTF}_i} : S^i \times K \times R_{\mathrm{OTF}} \times \prod_{j \in M} R_{\mathrm{SP}} \to \mathbb{R} \tag{6}$$

with

$$\begin{aligned}\pi_{\mathrm{OTF}_i}&\left(S_\ell^i, k_\ell^i, r_{\mathrm{OTF}_i}, r_{\mathrm{SP}}\right)\\ &= p_{S_\ell^i}\left(r_{\mathrm{OTF}_i}\right) - c(S_\ell^i, k_\ell^i) - \sum_{j \in M_{k_\ell^i}} p_{s_{ij}}\left(r_{\mathrm{SP}_j}\right)\end{aligned} \tag{7}$$

and his objective is

$$\max_{S_\ell^i \in S^i,\, k_\ell^i \in K^i} \quad \pi_{\mathrm{OTF}_i}\left(S_\ell^i, k_\ell^i, r_{\mathrm{OTF}_i}, r_{\mathrm{SP}}\right). \tag{8}$$

If there are several heterogeneous users, then OTF provider i's profit maximization problem needs to be considered over all users, to which he sells his composed services.

Summing up, the reputation information provided for elementary as well as composed services influences the OTF provider's economic decision problem via market prices driven by the users' demands.

In addition, within the service provider selection process, an OTF provider has to keep in mind that after he delivered the composed service to a user, he will receive a rating indicating the user's satisfaction. If the user's expectations are not met and the rating is negative, then this influences the future sale opportunities of an OTF provider. The incorporation of future profits makes the economic decision problem of an OTF provider even more complex. The profit maximization needs to be considered over current and future sales. Hence, the OTF provider's profit should be described as a discounted sum of profits that are expected in the long run from repeated interaction in the market.

C. Economic Decisions of Service Providers

The *Service Providers'* main challenge is to offer their elementary services such that an OTF provider decides to use these services for a composed service (*Service Provider Selection* in Figure 18). From a service provider's point of view, the difficulty is to provide services such that they match with those of other service providers technically as well as qualitatively. For a detailed description of the *Service Matching Process*, we refer to Section VII. As there is a huge number of elementary services available in the market, the service providers face an intensive competitive pressure. However, the services of different service providers are typically not perfectly exchangeable and, consequently, the OTF provider's decision for an elementary service depends on the availability of complementary services used for a composed service.

A service provider's profit in the OTF Computing market consists of the price he receives for his elementary service minus the costs to produce it. An economic decision of a service provider is the quality of the elementary service he delivers to an OTF provider. Suppose the set of possible quality levels is given by Q_{SP}. The profit of service provider $j \in M$ is

$$\pi_{\mathrm{SP_j}} : Q_{\mathrm{SP}} \times R_{\mathrm{OTF}} \times R_{\mathrm{SP}} \to \mathbb{R} \tag{9}$$

with

$$\pi_{\mathrm{SP_j}}\left(q_{ij}, r_{\mathrm{OTF}_i}, r_{\mathrm{SP}}\right) = p_{s_{ij}}\left(r_{\mathrm{SP}_j}\right) - c(q_{ij}). \tag{10}$$

The service provider's decision problem is to choose a quality level $q_{ij} \in Q_{\mathrm{SP}}$ that maximizes his profits,

$$\max_{q_{ij} \in Q_{\mathrm{SP}}} \pi_{\mathrm{SP_j}}\left(q_{ij}, r_{\mathrm{OTF}_i}, r_{\mathrm{SP}}\right). \tag{11}$$

Similarly to the OTF providers, a service provider maximizes his long run profits from repeated interaction in the market. The quality choice of a service provider has an impact on the quality of the composed service. After the execution of the composed service the OTF provider receives a rating and this rating influences the reputation values of the services and service providers used for the composition. Thus, a negative rating for a composed service may have the effect that the price a service provider receives from an OTF provider decreases. Accordingly, a service provider has to take the long run consequences of his short run decisions into account for his profit maximization.

D. Interaction in the OTF Computing Market

The interaction in the OTF Computing market can be formally described by means of game theoretic models. We start to explain the short run interaction for given current reputation values $r_{\mathrm{OTF}_i} \in R_{\mathrm{OTF}}$ and $r_{\mathrm{SP}_j} \in R_{\mathrm{SP}}$ for all $j \in M$.

If the short run decisions of users, OTF provider, and service providers are considered to be simultaneous, a *non-cooperative normal form game* can be used to describe the short run interaction. The players are users, OTF providers, and service providers. The strategic decision of a user is the selection of an OTF provider. An OTF provider selects the composed services he offers and the according composition strategies, while a service provider chooses the quality of the elementary services he is asked to deliver to the OTF provider. The payoffs of the market participants are the utility for a user and the profits for OTF and service providers derived in the previous subsections.

A well-known solution concept for non-cooperative normal form games is the *Nash equilibrium* [49]. In a Nash equilibrium, no market participant is able to increase his payoff by deviating from his strategic choice given the others' strategic choices. This means that in a Nash equilibrium

- a user cannot increase his utility by choosing a different OTF provider given the OTF providers' choices on composed services and composition strategy and given the service providers' quality choices,
- an OTF provider cannot increase his profit by changing the composed service or the composition strategy given the user's selection of an OTF provider and given the service providers' quality decisions,
- and a service provider cannot increase his profit by delivering his elementary service in a different quality given the user's selection of an OTF provider and given the OTF providers' choices on composed services and composition strategy.

We illustrate the normal form game by means of our running example. Consider the market of image processing services. Suppose there is one user who approaches an OTF provider to post-process his video. The OTF provider finds an elementary service that, however, from his point of view does not perfectly match the user's request (cf. Section VII). On the one hand, this matching result is imprecise as it is still dependent on the service provider's effort put into the performance of the video post-processing service. On the other hand, during the service composition process (cf. Section VIII), the OTF provider may put additional own effort to improve the quality of the video processing service. This example fits into the previously described economic framework as follows: There is one user, one OTF provider, and one service provider in the OTF Computing market. In addition, there is only one composed service available. As there is exactly one OTF provider who receives the user's request and there exists one composed service, the strategy set of the user is a singleton and can be described by $S = \{S_1\}$. The OTF provider has two different composition strategies $K = \{k_L, k_H\}$ (or different effort levels) and there are two different quality levels the service provider may choose for the elementary service $Q_{\mathrm{SP}} = \{q_L, q_H\}$. The qualities of the composed service are assumed to be $Q = \{LL, LH, HL, HH\}$. In addition, we take $u(S_1, LL) = 9$, $u(S_1, LH) = u(S_1, HL) = 10$, $u(S_1, HH) = 12$, $p_{S_1}(r_{\mathrm{OTF}_i}) = 8$, $c(S_1, k_L) = 1$, $c(S_1, k_H) = 2$, $p_{s_{11}}(r_{\mathrm{SP}_1}) = 4$, $c(q_L) = 2$, and $c(q_H) = 3$. The payoffs are shown in Figure 19. The entries of the payoff vectors are ordered such that the first value corresponds to the user's utility, the second to the OTF provider's profit and the third to the service provider's profit. The OTF provider chooses the row of the payoff matrix in Figure 19, the service provider the column and the user the matrix. The Nash equilibrium (S_1, k_L, q_L) with payoffs $(1, 3, 2)$ is marked in red.

S_1		q_L	q_H
	k_L	$(1, 3, 2)$	$(2, 3, 1)$
	k_H	$(2, 2, 2)$	$(4, 2, 1)$

Figure 19. Simultaneous Short Run Interaction

In this example, the user has no other option than to buy the composed service the OTF provider offers. Even if the user is interested in buying a composed service of high quality, OTF and service provider are willing to sell low quality as this is less expensive for them. In a more complicated scenario, the user may send a request to a different OTF provider or punish the OTF provider with a negative rating, which may lower the sale price in the long run.

Alternatively, the short run decisions of the market participants can be considered to be sequentially. In this case, a *non-cooperative extensive form game* should be used to describe the interaction in the OTF Computing market. The strategic choices are now sequential: First, the users select the OTF providers. Then, after receiving the users' requests, the OTF providers react and select composed services and composition strategies. Finally, the service providers choose the quality of the elementary services. The final payoffs are as previously described. A solution concept used for sequential decisions is the *subgame-perfect Nash equilibrium*, which is a refinement of the original Nash equilibrium for sequential interaction [50][51].

Figure 20 shows the previous example in an extensive form. The subgame perfect Nash equilibria can be obtained in our case by *backwards induction*. We first determine the best choices of the service provider who makes the last decision. Taking this decision as given, we determine the optimal choice of the OTF provider and finally the user decides. The optimal choices are market in red in Figure 20. The subgame perfect Nash equilibrium is (S_1, k_L, q_L) with payoffs $(1, 2, 3)$. Independently of the choice of the OTF provider, it is the best option for the service provider to produce an elementary service of low quality. The OTF provider now takes this choice of the service provider as given and deduces that his optimal action is to use composition strategy k_L.

Up to now, the prices in the OTF Computing market were assumed to be given depending on the reputation values of the OTF and service providers. There may exist a rule governing the evolution of prices depending on the strategic choices of

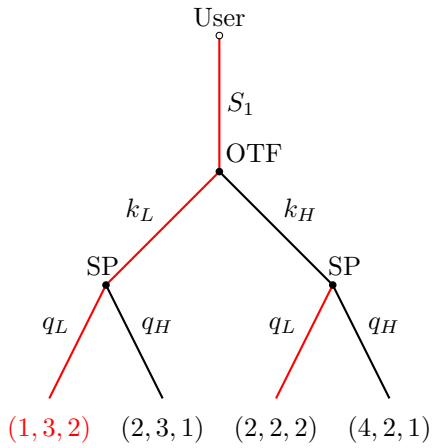


Figure 20. Sequential Short Run Interaction

the market participants and the rating finally received from the users.

X. Future Research Challenges

The introduction of a reputation system in the OTF Computing in Section V is conceptual and provides flexibility for further specifications. The research challenges resulting from the requirements imposed in Section IV have been highlighted in our previous contribution [1] and were investigated in more detail in Sections VII, VIII, and IX of this contribution. This section is devoted to further describe the trade-offs and challenges that are planed to be analyzed more intensively in future work.

Fuzzy Matching of Reputation Values: In Section VII, we analyzed in more detail how reputation should be matched. Since the reputation of a service is not an objective measure, such as signatures or protocols, uncertainty might be introduced into the matching process. For example, as noted in our fuzzy matching survey [12], the user stating the request might tolerate variations (e.g., "I want a service with *approximately* five stars"), or the request might include requirements, for which the corresponding information on the provider side do not exist yet (e.g., there has not been much feedback yet because the service is new on the market and thus the reputation is unclear). Section VII illustrated our fuzzy matching approach for a specific form of reputation values. However, there are still open issues in this area. For example, restrictions based on expert ratings (*R7*) could be introduced. Furthermore, we plan to quantify fuzziness based on necessity and possibility measures [52] in order to give even more feedback on the matching result.

Efficiency in Learning versus Privacy Protection: In Section VIII, we explained how a delayed feedback affects the convergence behavior of the learning process. The reinforcement learning approach, which is used to improve the quality of composed services, originally incorporates feedback immediately after a composition process. If the feedback is absent, the learning process is hampered. However, for reasons of privacy protection, no direct feedback is given to any party. Only an aggregated value of the accumulation of several individual ratings (feedback) is provided, as described in Section V. How the results of ambiguous feedback have to be interpreted in a more general context beyond the one described in Section VIII is still an open question. The influence of the execution order has to be thoroughly investigated in a different experiment in order to estimate, in which context a learning process benefits and suffers from ambiguous feedback, respectively. The presented results, however, are promising and gave a clear direction for our ongoing research. Furthermore, a combination of OTF provider-related ratings and publicly accessible ratings might be a good mean to improve the ratio between unambiguous feedback and ambiguous feedback. For example, an OTF provider might incorporate OTF provider-related rating values when available. When not available, ambiguous feedback can be incorporated. To measure the degree of ambiguity in order to enable an OTF provider to decide whether to use such feedback or not can be provided by the reputation value in terms of common measures of dispersion such as standard deviations or associated confidence intervals. In addition to these investigations, incorporating varying user ratings due to varying preferences is still an open challenge as well.

Economic Decisions: In Section IX, we described the economic decisions on the OTF Computing market and studied their interaction. Alternatively to an exogenously fixed rule of price evolution, the prices may also be strategically chosen by the OTF and service providers. Two possibilities are *price* (or *Bertrand*) and *quantity* (or *Cournot*) competition. In the first case, the providers compete by choosing profit maximizing prices and, in the second case, the providers compete by announcing prices such that the demanded quantities are profit maximizing. These two forms of competition are compared in [53], for example. The model in [53] allows for quality differences between the services and uses a parameter to describe substitutabilities or complementarities between the services. Within the model, the main observation is that for complementary services the providers' profits are higher if the providers compete in prices whereas for substitutable services with small quality differences quantity competition yields higher provider profits [53, Proposition 2]. However, if the services are substitutes and the quality differences are large, the provider with the quality advantage earns higher profits in case of price competition. An interesting direction for future research is to apply these models of strategic pricing to the OTF Computing market including a reputation system.

More complex pricing structures such as *two-part tariffs* with a fixed fee and usage depended price or *three-part tariffs* with an additionally included volume have been analyzed in [54] applied to a cloud computing context. The main observation is that in case of symmetric providers and homogeneous services the competitive pressure forces the equilibrium prices to decrease until they are equal to the providers' costs. Hence, for the OTF Computing market, on which typically both substitutable and complementary services are traded and pricing structures may be even more complex, further investigations are needed.

A different point that our simple example already indicates is that optimal decisions in the short run may not coincide with those in the long run when the impact of the reputation information in the market prices is taken into account. Therefore, the long run interaction in the market have to be further

investigated. The simultaneous or sequential short run game in the OTF Computing market has to be considered as repeated game or as a stochastic game [55]. Without a reputation system and from a contract design perspective, the long run decisions in the OTF Computing market are analyzed in [56]. It is shown, that the OTF provider's evaluation of future profits plays a crucial role for the long run equilibrium quality level in the market for composed services. If the OTF provider's future expectations are too pessimistic, then composed services of low quality are going to be offered in the OTF Computing market even when high quality is more efficient from a welfare perspective. In [57], simulation techniques for complex strategic decisions are applied to the OTF Computing market. However, further analyzes to investigate the long run decisions in the market for composed services and to understand the impact of reputation information on the market participants' strategic behavior are necessary. This is one main direction of our current and future research.

Benefit of Privacy Protection: As discussed in this paper and [1], the design of a privacy-preserving solution entails a multitude of trade-offs that need to be taken into account, e.g., the trade-off between privacy and learning mechanism efficiency. Thus, it needs to be investigated whether market participants are interested in implementing a privacy-preserving solution at all. We need to prove that privacy protection is a benefit of OTF Computing and that users rather use such a market than any other, which does not provide such strong privacy guarantees. Concerning the introduced reputation system, we want to examine whether users are more willingly providing ratings when their privacy is protected—which is not the case in any other state-of-the-art reputation system in use today.

Manipulation Resistance versus Privacy Protection: An important further issue is to obtain truthful user feedback. Ratings may be dishonest or randomly chosen [23]. So far we assumed that users have no incentives to strategically manipulate their feedback and moreover we supposed that feedback on a transaction is always provided. Truthful rating behavior is induced by *incentive compatible reputation mechanisms* [58] (and the references mentioned therein). To ensure privacy protection, several ratings need to be accumulated and aggregated. It has already been analyzed how the aggregation of ratings impacts the efficiency of a reputation mechanism [59] and how it influences incentives for truthful rating behavior [60]. An important next step now is to further understand the interplay of incentive compatibility and privacy protection. Therefore, a challenging question is whether and how it is possible to design reputation systems that induce truthful feedback and respect privacy protection.

XI. Conclusion

In the context of OTF Computing, we interpret reputation information as a signal for quality in markets of composed services. From an economic perspective, the buying decision of a user and the future sale opportunity of an OTF provider crucially depend on the current reputation value. For that reason, we proposed a conceptual design of a reputation system that collects information about experiences users make with composed services in transactions. We identified and described requirements for such a reputation system from a technical, and economic perspective, and presented research challenges that automatically emerge from conflicting objectives. In this regard, we focused on three aspects, which are among others crucial for markets of composed services: service matching, service composition and economic decisions of market participants (users, OTF providers, and service providers).

In case of service matching, we introduced our fuzzy matching approach that is able to cope with reputation values that are – in comparison to other non-functional properties such as performance or costs – not of objective nature. Nevertheless, there are still open challenges such as the integration of restrictions based on expert ratings. Furthermore, we intend to quantify fuzziness based on necessity and possibility measures in order to give even more feedback on the matching result.

The impact of reputation information for our RL based composition and recommendation approach has to be investigated in a more general context. For our image processing example with particular characteristics, however, we presented promising results. A major challenge in this context is the collection and processing of user ratings from different sources (e.g., OTF provider-related vs. publicly available) in order to increase the amount of learning samples in terms of reputation information.

From the economic point of view, we analyzed the individual decision problems of the market participants in the OTF Computing market as well as their interaction. Hereby, we explicitly included the information provided by the reputation system. Moreover, we outlined by means of a simple example how strategic decisions may influence the quality of composed services traded in equilibrium in the market. Some preliminary results to analyze the economic interaction in the OTF Computing market have already been obtained in previous and ongoing work as outlined in Section X. This can be seen as a first step for a comprehensive analysis of the impact of reputation on the OTF Computing market.

The contribution of this paper is not necessarily restricted to OTF Computing alone. Results of our work can also be adopted to other areas, in which reputation of combinable products is crucial. For our future work, our reputation system can be considered as an interface to bring together approaches and results from different (sub-)disciplines such as economics, software engineering, distributed systems, artificial intelligence, and security.

Acknowledgment

This work was partially supported by the German Research Foundation (DFG) within the Collaborative Research Center "On-The-Fly Computing" (SFB 901).

References

[1] S. Brangewitz, A. Jungmann, R. Petrlic, and M. C. Platenius, "Towards a flexible and privacy-preserving reputation system for markets of composed services," in Proceedings of the Sixth International Conferences on Advanced Service Computing (SERVICE COMPUTATION), 2014, pp. 49–57.

[2] R. Petrlic, A. Jungmann, M. C. Platenius, W. Schaefer, and C. Sorge, "Security and privacy challenges in on-the-fly computing," in Tagungsband der 4. Konferenz Software-Technologien und -Prozesse (STeP), 2014, pp. 131–142.

[3] "Collaborative Research Center 901 - On-The-Fly Computing," 2014, URL: http://sfb901.uni-paderborn.de [accessed: 2014-11-28].

[4] "Instagram," 2014, URL: http://www.instagram.com [accessed: 2014-11-28].

[5] F. Gómez Mármol and M. Q. Kuhnen, "Reputation-based web service orchestration in cloud computing: A survey," Concurrency and Computation: Practice and Experience, 2013.

[6] N. Hiratsuka, F. Ishikawa, and S. Honiden, "Service selection with combinational use of functionally-equivalent services," in Proceedings of the 18th IEEE International Conference on Web Services (ICWS), 2011, pp. 97–104.

[7] J. Peer, "Web service composition as ai planning - a survey," University of St. Gallen, Switzerland, Tech. Rep., 2005.

[8] P. Bartalos and M. Bieliková, "Semantic web service composition framework based on parallel processing," in Proceedings of the 11th IEEE Conference on Commerce and Enterprise Computing (CEC), 2009, pp. 495–498.

[9] P. Bartalos and M. Bieliková, "Automatic dynamic web service composition: A survey and problem formalization," Computing and Informatics, vol. 30, no. 4, 2011, pp. 793–827.

[10] M. Aiello, E. el Khoury, A. Lazovik, and P. Ratelband, "Optimal QoS-Aware Web Service Composition," in Proceedings of the 11th IEEE Conference on Commerce and Enterprise Computing (CEC), 2009, pp. 491–494.

[11] M. C. Platenius, "Fuzzy service matching in on-the-fly computing," in Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering (ESEC/FSE). ACM, 2013, pp. 715–718.

[12] M. C. Platenius, M. von Detten, S. Becker, W. Schäfer, and G. Engels, "A survey of fuzzy service matching approaches in the context of on-the-fly computing," in Proceedings of the 16th International ACM Sigsoft Symposium on Component-based Software Engineering (CBSE). ACM, 2013, pp. 143–152.

[13] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction. Cambridge, Massachusetts: MIT Press, 1998.

[14] A. Jungmann and B. Kleinjohann, "Learning Recommendation System for Automated Service Composition," in Proceedings of the 10th IEEE International Conference on Services Computing (SCC), 2013, pp. 97–104.

[15] A. Jungmann, B. Kleinjohann, and L. Kleinjohann, "Learning service recommendations," Int. J. Business Process Integration and Management, vol. 6, no. 4, 2013, pp. 284–297.

[16] A. Van Lamsweerde, "Goal-oriented requirements engineering: A guided tour," in Proceedings of the Fifth IEEE International Symposium on Requirements Engineering (RE), 2001, pp. 249–262.

[17] Y. Wang and J. Vassileva, "A review on trust and reputation for web service selection," in 27th International Conference on Distributed Computing Systems Workshops (ICDCSW), 2007, pp. 25–25.

[18] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in Proceedings of the IEEE Symposium on Security and Privacy (SP), 2008, pp. 111–125.

[19] R. Petrlic, S. Lutters, and C. Sorge, "Privacy-preserving reputation management," in Proceedings of the 29th Symposium On Applied Computing. ACM, 2014.

[20] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in Proceedings of the 17th international conference on Theory and application of cryptographic techniques. Springer, 1999, pp. 223–238.

[21] C. Shapiro, "Premiums for high quality products as returns to reputations," The Quarterly Journal of Economics, vol. 98, no. 4, 1983, pp. 659–680.

[22] H. Bar-Isaac and S. Tadelis, "Seller reputation," Foundations and Trends in Microeconomics, vol. 4, no. 4, 2008, pp. 273–351.

[23] E. Friedman, P. Resnick, and R. Sami, "Manipulation-resistant reputation systems," in Algorithmic Game Theory, Chapter 27. Cambridge University Press, 2007.

[24] C. Dellarocas, "Reputation mechanism design in online trading environments with pure moral hazard," Information Systems Research, vol. 16, no. 2, 2005, pp. 209–230.

[25] F. Gómez Mármol and G. Martinez Pérez, "Trust and reputation models comparison," Internet Research, vol. 21, no. 2, 2011, pp. 138–153.

[26] F. Kerschbaum, "A verifiable, centralized, coercion-free reputation system," in Proceedings of the 8th ACM workshop on Privacy in the electronic society (WPES), 2009, pp. 61–70.

[27] E. Androulaki, S. G. Choi, S. M. Bellovin, and T. Malkin, "Reputation systems for anonymous networks," in Proceedings of the 8th International Symposium on Privacy Enhancing Technologies. Springer, 2008, pp. 202–218.

[28] A. Shaikh Ali, S. Majithia, O. F. Rana, and D. W. Walker, "Reputation-based semantic service discovery," Concurrency and Computation: Practice and Experience, vol. 18, no. 8, 2006, pp. 817–826.

[29] M. R. Motallebi, F. Ishikawa, and S. Honiden, "Component trust for web service compositions," in AAAI Spring Symposium Series, 2012.

[30] Z. Malik and A. Bouguettaya, "Rateweb: Reputation assessment for trust establishment among web services," The VLDB Journal, vol. 18, no. 4, 2009, pp. 885–911.

[31] K. Huang, J. Yao, Y. Fan, W. Tan, S. Nepal, Y. Ni, and S. Chen, "Mirror, mirror, on the web, which is the most reputable service of them all?" in Service-Oriented Computing. Springer, 2013, pp. 343–357.

[32] S.-E. Tbahriti, M. Mrissa, B. Medjahed, C. Ghedira, M. Barhamgi, and J. Fayn, "Privacy-aware daas services composition," in Database and Expert Systems Applications, 2011, pp. 202–216.

[33] E. Costante, F. Paci, and N. Zannone, "Privacy-aware web service composition and ranking," in Proceedings of the 20th IEEE International Conference on Web Services (ICWS), 2013, pp. 131–138.

[34] M. C. Platenius, S. Arifulina, R. Petrlic, and W. Schäfer, "Matching of incomplete service specifications exemplified by privacy policy matching," in 4th International Workshop on Adaptive Services for the Future Internet. Springer, 2014, in press.

[35] L. A. Zadeh, "Fuzzy sets," Information and control, vol. 8, no. 3, 1965, pp. 338–353.

[36] H. Lam, F. H. F. Leung, and P. K.-S. Tam, "Stable and robust fuzzy control for uncertain nonlinear systems," Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, vol. 30, no. 6, 2000, pp. 825–840.

[37] R.-E. Precup, M. L. Tomescu, and S. Preitl, "Fuzzy logic control system stability analysis based on lyapunovs direct method," International Journal of Computers, Communications & Control, vol. 4, no. 4, 2009, pp. 415–426.

[38] Y. Yi, T. Fober, and E. Hüllermeier, "Fuzzy operator trees for modeling rating functions," International Journal of Computational Intelligence and Applications, vol. 8, no. 04, 2009, pp. 413–428.

[39] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 3rd ed. Upper Saddle River, NJ, USA: Prentice Hall, 2009.

[40] J. Rao and X. Su, "A survey of automated web service composition methods," in Proceedings of the First International Conference on Semantic Web Services and Web Process Composition (SWSWPC). Springer-Verlag, 2005, pp. 43–54.

[41] M. Ghallab, D. Nau, and P. Traverso, Automated planning: theory & practice. San Francisco, CA, USA: Morgan Kaufmann, 2004.

[42] D. Doliwa, W. Horzelski, M. Jarocki, A. Niewiadomski, W. Penczek, A. Pólrola, M. Szreter, and A. Zbrzezny, "Planics - a web service composition toolset," Fundam. Inf., vol. 112, no. 1, 2011, pp. 47–71.

[43] A. Jungmann, F. Mohr, and B. Kleinjohann, "Combining automatic service composition with adaptive service recommendation for dynamic markets of services," in Proceedings of the IEEE 10th World Congress on Services (SERVICES), 2014, pp. 346–353.

[44] M. L. Puterman, Markov decision processes: Discrete stochastic dynamic programming. Hoboken, NJ, USA: Wiley-Interscience, 2005.

[45] C. J. C. H. Watkins and P. Dayan, "Q-learning," Machine Learning, 1992, pp. 279–292.

[46] T. M. Mitchell, Machine Learning. New York: McGraw-Hill, 1997.

[47] A. Mas-Colell, M. D. Whinston, and J. R. Green, Microeconomic Theory. Oxford University Press, 1995.

[48] A. M. Spence, "Monopoly, quality, and regulation," The Bell Journal of Economics, vol. 6, no. 2, 1975, pp. 417–429.

[49] J. Nash, "Non-cooperative games," Annals of Mathematics, vol. 54, no. 2, 1951, pp. 286–295.

[50] R. Selten, "Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit: Teil I: Bestimmung des dynamischen Preisgleichgewichts," Zeitschrift für die gesamte Staatswissenschaft, vol. 121, no. 2, 1965, pp. 301–324.

[51] R. Selten, "Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit: Teil II: Eigenschaften des dynamischen Preisgleichgewichts," Zeitschrift für die gesamte Staatswissenschaft, vol. 121, no. 4, 1965, pp. 667–689.

[52] D. Dubois, H. Nguyen, and H. Prade, "Possibility theory, probability and fuzzy sets misunderstandings, bridges and gaps," in Fundamentals of Fuzzy Sets, ser. The Handbooks of Fuzzy Sets Series. Springer US, 2000, vol. 7, pp. 343–438.

[53] J. Häckner, "A note on price and quantity competition in differentiated oligopolies," Journal of Economic Theory, vol. 93, no. 2, 2000, pp. 233–239.

[54] J. Künsemöller, S. Brangewitz, H. Karl, and C.-J. Haake, "Provider competition in infrastructure-as-a-service," in Proceedings of the IEEE International Conference on Services Computing (SCC), 2014, pp. 203–210.

[55] G. J. Mailath and L. Samuelson, Repeated games and reputations: long-run relationships. Oxford university press, 2006.

[56] S. Brangewitz, C.-J. Haake, and J. Manegold, "Contract design for composed services in a cloud computing environment," in Proceedings of the 2nd International Workshop on Cloud Service Brokerage (CSB), 2014, in press.

[57] M. Feldotto and A. Skopalik, "A simulation framework for analyzing complex infinitely repeated games," in Proceedings of the 4th International Conference on Simulation and Modeling Methodologies, Technologies and Applications (SIMULTECH). SciTePress, 2014, pp. 625–630.

[58] S. Phoomvuthisarn, "A survey study on reputation-based trust mechanisms in service-oriented computing," Journal of Information Science and Technology, vol. 2, no. 2, 2011, pp. 1–12.

[59] C. Dellarocas, "How often should reputation mechanisms update a trader's reputation profile?" Information Systems Research, vol. 17, no. 3, 2006, pp. 271–285.

[60] C. Aperjis and R. Johari, "Optimal windows for aggregating ratings in electronic marketplaces," Management Science, vol. 56, no. 5, 2010, pp. 864–880.

DAiSI—Dynamic Adaptive System Infrastructure: Component Model and Decentralized Configuration Mechanism

Holger Klus
ROSEN Technology & Research Center GmbH
Lingen (Ems), Germany
hklus@rosen-group.com

Andreas Rausch, Dirk Herrling
Technische Universität Clausthal
Clausthal-Zellerfeld, Germany
{andreas.rausch, dirk.herrling}@tu-clausthal.de

Abstract— **Dynamic adaptive systems are systems that change their behavior according to the needs of the user at run time, based on context information. Since it is not feasible to develop these systems from scratch every time, a component model enabling dynamic adaptive systems is called for. Moreover, an infrastructure is required that is capable of wiring dynamic adaptive systems from a set of components in order to provide a dynamic and adaptive behavior to the user. In this paper we present just such an infrastructure or framework—called Dynamic Adaptive System Infrastructure (DAiSI). Because DAiSI has been developed for a number of years, we will cover as well the history of DAiSI as the newest advances. We will present an example illustrating the adaptation capabilities of the framework we introduce. The focus of the paper is on the underlying component model of DAiSI and the decentralized configuration mechanism.**

Keywords— dynamic adaptive systems; component model; component composition; adaptation; componentware; component container; decentralized configuration.

I. INTRODUCTION

This paper is an extended version of a paper presented at the Sixth International Conference on Adaptive and Self-Adaptive Systems and Applications [1].

Software-based systems pervade our daily life—at work as well as at home. Public administration or enterprise organizations can scarcely be managed without software-based systems. We come across devices executing software in nearly every household. The continuous increase in size and functionality of software systems has made some of them among the most complex man-made systems ever devised [2].

In the last two decades the trend towards "everything, every time, everywhere" has been dramatically increased through a) smaller mobile devices with higher computation and communication capabilities, b) ubiquitous availability of the Internet (almost all devices are connected with the Internet and thereby connected with each other), and c) devices equipped with more and more connected, intelligent and sophisticated sensors and actuators.

Nowadays, these devices are increasingly used within an organically grown, heterogeneous, and dynamic IT environment. Users expect them not only to provide their primary services but also to collaborate autonomously with each other and thus to provide real added value. The challenge is therefore to provide software systems that are robust in the presence of increasing challenges such as change and complexity [3].

The reasons for the steady increase in complexity are twofold: On the one hand, the set of requirements imposed on software systems is becoming larger and larger as the extrinsic complexity increases, in the form of, for example, additional functionality and variability. In addition, the structures of software systems—in terms of size, scope, distribution and networking of the system among other things—are themselves becoming more complex, which leads to an increase in the intrinsic complexity of the system.

Change is inherent, both in the changing needs of users and in the changes, which take place in the operational environment of the system. Hence, it is essential that our systems are able to adapt to maintain the satisfaction of the user expectations and environmental changes in terms of an evolutionary change. Dynamic change, in contrast to evolutionary change, occurs while the system is operational. Dynamic change requires that the system adapts at run time.

Since the complexity and change may not permit human intervention, we must plan for automated management of adaptation. The systems themselves must be capable of determining what system change is required, and in initiating and managing the change process wherever possible. This is the aim of self-managed systems.

Self-managed systems are those capable of adapting to the current context as required though self-configuration, self-healing, self-monitoring, self-tuning, and so on. These are also referred to as self-x, autonomic systems. Additionally, new components may enter or leave the system at run time. We call those systems 'dynamic adaptive'.

Providing dynamic adaptive systems is a great challenge in software engineering [3]. In order to provide dynamic adaptive systems, the activities of classical development approaches have to be partially or completely moved from development time to run time. For instance, devices and software components can be attached to a dynamic adaptive system at any time. Consequently, devices and software components can be removed from the dynamic adaptive system or they can fail as the result of a defect. Hence, for dynamic adaptive systems, system integration takes place during run time.

To support the development of dynamic adaptive systems a couple of infrastructures and frameworks have been developed, as discussed in a related work section, Section II. In our research group we have also developed a framework for dynamic adaptive (and distributed) systems, called DAiSI.

We believe that service oriented, component based, dynamic adaptive systems need to address at least three kinds of adaptation:

1. Component service implementation adaptation: The implementation of a service is changed within the component that provides it. Thus, only the output of the component is affected but not the overall structure of the application. A simple example is the sorting order of search results that can be modified by the end user.
2. Component service usage adaptation: A component that uses a service of another component switches the service provider, i.e., another component that provides the same service. This can happen, for example, because the service quality of the now used service is superior compared to the formerly used one. Component service usage adaptation may yield to a better service quality of the component switching services. Only the directly involved components are affected by the change in system configuration.
3. System configuration adaptation: If the provided services of a component change, these changes can cascade through the complete system because optional dependencies could be resolved or mandatory dependencies can no longer be resolved. We speak of system configuration adaptation in such cases.

The development of the DAiSI was always motivated through running application examples and demonstrators. Based on the evaluation results a couple of drawbacks were identified. I) DAiSI's component model was not able to handle service cardinalities, such as exclusive and shared use of a specific service or service reference sets. Most of the applications realized needed service cardinalities. Due to the absence of service cardinalities we had to create workarounds. II) DAiSI's dynamic configuration mechanism was realized as a centralized component. The centralized configuration component was easy to implement but obviously it turned out to be a bottleneck.

For those reasons we have developed and implemented an improved version of the DAiSI framework. It contains a sophisticated component model including service cardinalities and a decentralized system configuration mechanism. In this paper, the new version of the DAiSI framework will be presented.

The rest of the paper is structured as follows: After a short description of the related work we provide an overview of the DAiSI framework in Section III. DAiSI's main essential, a domain model, an adaptive component model, and a decentralized dynamic configuration mechanism are introduced in this section. Then we describe a small sample application to illustrate the decentralized dynamic configuration mechanism of the adaptive components in Section IV. A short conclusion will round the paper up.

II. Related Work

Component-based software development, component models and component frameworks provide a solid approach to support evolutionary changes to systems. It is a well understood method that proved useful in numerous applications. Components are the units of deployment and integration. This allows high flexibility and easy maintenance. During design time components may be added or removed from a system [4].

However, the early component models did not provide means of adding or removing components from a running system. Also, the integration of new interaction links (e.g., component bindings) was not possible. Service-oriented approaches stepped up to the challenge. These systems usually maintain a service repository, in which every component that enters the system is registered. A component that wants to use such a component can query the service register for a matching service and connect to it, if one is found. For the domain of dynamic systems this means that a component can register its provided and required services. If a suitable service provider for one of the required services registers itself, it can be bound to satisfy the required service [5].

Service-oriented approaches have the inconvenient characteristic of not dealing with the adaptability of components. A component developer is solely responsible for the implementation of the adaptive behavior. This starts at the application logic and stretches to the discovery of unresponsive services, the discovery of newly available service, the discovery of services with a better quality of service, and so on. A couple of frameworks have been developed to support dynamic adaptive behavior, while, at the same time, making it easier for the developer to focus on implementing the behavioral changes in his component.

One of the first frameworks to support dynamic adaptive components was CONIC. It defines a description language that can be used to change the structure of modules of a running system. It allows the termination of running and the instantiation of new modules and the creation or deletion of links between them. CONIC was controlled through a centralized management console where the procedures for the reconfiguration could be entered [6].

REX is another framework for the support of dynamic-adaptive systems. It used the experience gained in the research for CONIC and aimed at dynamic adaptive, parallel, distributed systems. The concept was that such systems consist of components that are linked by interfaces. A new interface description language was invented, to be able to describe the interfaces. Components were seen as types, allowing multiple instances of every component to be present at run-time. Just like CONIC, REX allowed the creation and termination of component instances and the links between them. Both, CONIC and REX share the disadvantage that they support dynamic reconfiguration only through explicit reconfiguration programs. These need to be different for every situation that is detected and intended.

The approach moves the adaptation logic out of the component, but nevertheless, the developer has to deal with the adaptation strategy for every possible occurring change [7], [8].

Current frameworks such as ProAdapt [9] and Config.NETServices [10] have a more generic adaption and configuration mechanism. Components that were not known during the design-time of the system can be added or removed from the dynamic adaptive system during run-time. Therefore, a generic component configuration mechanism is provided by the framework. As with our first version of the DAiSI framework, these frameworks are based on a centralized configuration mechanism. Moreover, the underlying component model is restricted—for instance the exclusive usage of services cannot be described.

These are the two main issues we address with this paper. On the one hand, a self-organizing infrastructure is needed to remove the centralized configuration service as a single point of failure. On the other hand, service cardinalities are called for. Many service should only be used exclusively (e.g., security relevant services) or might only be used by a certain number of service consumers (e.g., a component that is running on a node with limited computing power). In the following section we will therefore present the DAiSI approach that addresses these issues.

III. DAiSI – Dynamic Adaptive System Infrastructure

Our approach for self-organizing systems is based on a specific framework called DAiSI [11], [12], [13], [14]. DAiSI consists of three main parts or elements: a domain model, an adaptive component model, and a decentralized dynamic configuration mechanism. In the following section, we will cover the history of the DAiSI approach, which gives an overview of the underlying concepts and realized industry projects and prototypes.

A. History of the DAiSI Approach

The research towards DAiSI started in 2004 [15] and the first version was implemented and published in 2006/07 [11], [12], [13], [14]. Based on the DAiSI framework, a couple of dynamic adaptive systems (research and industrial demonstrators) were developed and evaluated. Some of them were developed into successful business applications, for example [16] and [17]. The demonstrators that have been built were summarized in [18] and are briefly sketched in the following paragraphs:

Assisted Bicycle Training: In 2005, we proposed an ambient intelligence system for the training of a cycling group [19]. The individual training of one cyclist in the group is optimized based on the readings of numerous sensors, which evaluate his physical condition. Based on this information and the physical condition of the other cyclists in the group an optimal position for every cyclist is calculated, which has an influence on the training mainly because of the slipstream. We published the results of the research regarding the simulation of a cycling group in [20]. DAiSI was used in this scenario to connect the cyclist among each other, as well as the cyclists with a team cycling trainer, if he belongs to the same team. The resulting demonstrator has been exhibited at the CeBIT fair in 2005.

Assisted Living: In the assisted living scenario the focus lay on the monitoring of elderly people – more specific on their food and beverage consumption and their overall health status. The research started in 2005 under the assumption that our society is aging continuously and the expenses for health care are steadily increasing. On the other hand, more and more people prefer to stay in their known environment when they are aging. To address this issue, we proposed an apartment equipped with a multitude of sensors and intelligent devices (e.g., a fridge that monitors its contents). All these devices monitor and evaluate the state of the elderly person (e.g., vital data, did the person fall, did he/she drink enough, etc.) and the apartment (e.g., if the food in the refrigerator is still edible or already spoilt). DAiSI was used in this scenario to connect all sensors and devices so that new components can be installed and removed at run time [21], [12], [22].

Assisted Cross Country Skiing: In 2008, at the CeBIT fair, the Sport Information System (SiS) was exhibited. It was targeted at the cross country skiing domain. The system allowed a skiing trainer to analyze the skiing technique of one or more cross country skiers. If no radio connection between the trainer and the trainee could be established, the DAiSI configured the system in a way that the trainee got feedback regarding his technique based on an automated analysis of the movement of his skiing sticks. The analysis was possible because the skiing sticks were equipped with sensor nodes and the trainee carried a personal digital assistant (PDA) with himself.

Emergency Management System: In 2009, the Emergency Management System was exhibited at the CeBIT fair. Its goal was to support rescue workers in a mass casualty incident, also known as major incident. In these incidents, the rescue workers are usually outnumbered by the amount of casualties. To be able to deal best with such incidents, a good overview of the incident site and the casualties is mandatory. The casualties' treatment needs to be prioritized to save as many lives as possible. The emergency management system includes sensor nodes for every casualty and tablet-like computers for medics. These devices allow the computation of an interactive map that displays all rescue workers and casualties. Additional special hardware allows the monitoring of vital data of casualties and therefore the automatic suggestion of treatment priorities [23], [24], [25], [26]. The system is highly dynamic, because casualties and medic are continuously entering or leaving the system.

SmartSchank: In 2008, the project launched under the name "HomeS". It was designed to reduce the loss of beverage in the gastronomy through drawn, but not sold drinks. As hardware components, the system featured intelligent beer taps and registers with a credit/debit system. One requirement was the installation of the system without the need to manually configure it. Additionally, it was supposed to work in small pubs as well as in large arenas. DAiSI was used as a conceptual platform that could fulfill

these requirements. In 2010, the prototype was exhibited at the CeBIT fair [16]. It was later evolved to a commercial product by the project partner [17].

Smart City and Smart Airport: In the context of the NTH Focused Research School for IT Ecosystems a demonstrator for a smart airport as an example for an IT ecosystem was built. All users within the smart airport were supposedly represented by devices called SmartFolks. These devices also served as an interface to the ecosystem in the airport. Two sub projects in this project were the Smart CheckIn and the RuleIT methodology. While the first scenario enables travelers to choose if they would rather pay less, but wait longer in a line for a standard check in, or if they would prefer to pay a little bit extra and get a guaranteed time slot to check in without needing to wait at all, the latter scenario focused on rule based application configuration and introduced user decisions into the application configuration process [14], [27], [28], [29].

Pac-Man: In the OPEN project (in 2010), the migration of an application at run time was researched. The focus lay on preserving the internal state of the application so that the end user can migrate his application from one device to the other. Additionally, the graphical user interface was supposed to adjust to the device the application was migrated to. The classic game Pac-Man was chosen as an application example, while DAiSI served as the underlying system infrastructure [30], [31].

Biathlon Training: This is the sample application that is used in this publication. The details will be introduced in Section IV.

We found that the three basic adaptation requirements that have been mentioned above, as well as some additional features can best be realized by a number of different architectural concepts:

Component Model: DAiSI was invented to support the development of service oriented, component based systems. Components communicate with each other through provided and required services. A subset of the required services can enable the component to provide a subset of the provided services. This relation is expressed by the so-called component configuration. The component model features all these elements since early development stages and will be further explained throughout the rest of this paper.

Configuration Service: The configuration service composes the application at run time of the present DAiSI components, by binding required and provided services to each other and configuring every component.

Registry Service: The configuration service requires knowledge of the present component in order to configure the system. The registry service maintains a database of all installed DAiSI components and the nodes they are deployed on. It monitors their responsiveness to discover and remove unresponsive components from the system.

Device Bay: Small scale devices do not always have the necessary computing power or memory to be able to participate directly in a DAiSI application. The device bay concept enables nodes with a higher computing power to represent those devices using an adapter pattern.

Dependability: In early DAiSI implementations provided services could satisfy required services if the interfaces describing the services were syntactically compatible. Further research enabled the requirement of semantic compatibility that can be ensured based on run time equivalence class testing.

Migrateability: Moving a running DAiSI component from one node to another was a requirement the DAiSI needed to satisfy for a particular research project. Through an extension of the component model, which allows the extraction and insertion of a consistent internal state, this requirement can be met. The migration is realized in three steps. At first the internal state of the component is extracted. In the second step, the component is stopped on the old node and started on the new node. Before it is activated, the internal state is inserted into the freshly started component in a third step.

Self-Organization: In early versions of the DAiSI, the system was configured by the configuration service, which followed an optimization algorithm and configured one component after the other. With large scale applications the configuration through one central configuration service becomes a bottleneck. Therefore, the configuration logic was moved into the individual components and they became self-organizing components. Details to the self-organization capabilities of the DAiSI can be found in Sections III and IV.

Architecture Awareness: Although syntactic and semantic compatibility already limit, which components can use which services of other components, this is often not enough. Dependent on the application domain, additional architectural rules may need to be enforced. For example, should a cross country skier only be linked to two skiing sticks; both need to belong to him. Just ensuring that he is connected to two skiing sticks does not fulfill the requirements of the end user. How the self-organizing components ensure that the application architecture requirements are met can be found in Sections III and IV.

Component Market: In conflicting cases where two or more components would like to use a particular service that cannot be used by all of them, a decision needs to be made. A component market solves this issue by introducing a currency into the system. Components that want to use service of a different component can use their currency to bid for a service. If they are chosen, they transfer their currency, and the component providing the service has both, the newly earned and its default currency. It can use that to bid for other services to improve its service quality. The underlying idea is that overall system quality improves if a component that has its provided services used by many other components improves its service quality.

User Decisions: The composition of an application out of the available components follows either an optimization goal (e.g., build the application that integrates the most components), or a set of rules. In different scenarios this can lead to more than one possible solution, which are to be considered as of the same quality. In these cases the end user can be involved to decide, which way the application should be configured.

Numerous publications discuss these architectural concepts with regard to a specific demonstrator or industry application. Table I shows a matrix with the different concepts in the top row and the demonstrators in the first column. The bibliography entries in the intersections indicate, which publication explains which concepts with the help of which demonstrators. In cases where a concept was used for a particular demonstrator but no results have been published, a "yes" is written in the matrix to state the fact.

TABLE I. MATRIX MAPPING ARCHITECTURAL CONCEPTS AND DEMONSTRATORS TO PUBLICATIONS

	Component Model	Configuration Service	Registry Service	Device Bay	Dependability	Migrateability	Self-Organization	Architecture Awareness	Component Market	User Decisions
Assisted Bicycle Training	[19]	[19]		[19]						
Assisted Living	[32] [12] [13] [33]	[32] [12] [13]	[13] [22]	[32] [12] [13]						
Assisted Cross Country Skiing	yes	yes	yes	yes						
Emergency Management System	[25] [35] [36] [26]	[25] [35] [36] [26] [34]			[25] [35] [36] [26]				[34]	
SmartSchank	yes	yes	yes							
Smart City / Smart Airport		[28] [29]								[29] [37]
Pac-Man	[38]					[38] [31]				
Biathlon Training	[39] [1]	[39] [1]					[39] [1]	[39] [1]		
None of the above	[11]	[11]	[11]							

B. Domain Model

The three elements of the DAiSI framework – the domain model, an adaptive component model, and a decentralized dynamic configuration mechanism will be introduced in this section. The three elements and their relationship to each other are depicted in Figure 1 using a UML class diagram. Note, a complete description of the DAiSI framework can be found in [39].

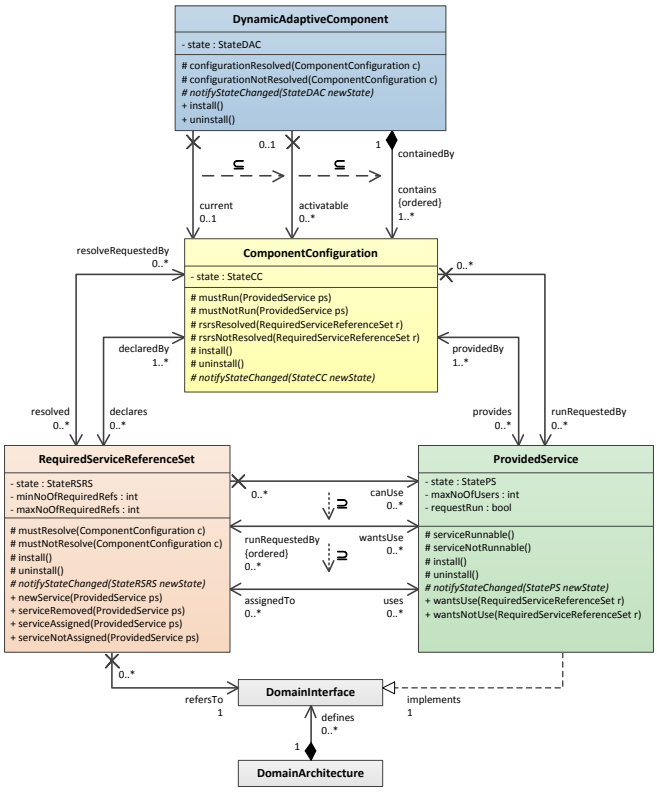


Figure 1. Core elements of the DAiSI framework.

As in other domains, such as the network domain, physical connectors (like the RJ 45 connector) and their pin configurations are standard and well known by all component vendors. A similar situation can be found in the operating system domain: The interface for printer drivers is standardized and published by the operating system vendor. Third-party printer vendors adhere to this interface specification to create printer drivers that are plugged into the operating system during run time.

The same principle is used in the DAiSI framework: The domain model contains standardized and broadly accepted interfaces in the domain. The domain model defines the basic notions and concepts of the domain shared by all components. This means the domain model provides the foundation for the dynamic configuration of the adaptive system and the available components.

The domain model, as shown in Figure 1, consists of the *DomainInterface* and *DomainArchitecture* classes. The domain model itself is represented by an instance of the *DomainArchitecture* class. A domain model contains a set of domain interfaces, represented by an instance of the class *DomainInterface*.

Domain interfaces contain syntactical information like method signatures or datatypes occuring in the interfaces. In addition they may also contain a behavioral specification of the interface following the design by contract approach, for

instance using pre- and postconditions and invariants to describe the functional behavior of a domain interface [25].

Usually, components need services from other components to provide their own service within the dynamic adaptive system. To indicate, which services a component provides and requires it refers to the corresponding *DomainInterface*. As components providing services and components requiring services refer to the same domain interface description DAiSI is able to identify those and bind these components together during run time.

Using simple domain interface descriptions the correctness of the binding can only be guaranteed on a syntactical level. Once the domain interface descriptions contain additional information about the functional behavior, the correctness of the binding can also be guaranteed on the behavioral level. Therefore, we have developed a sophisticated approach based on run-time testing. Further information of DAiSI's solution to guarantee functional correctness of dynamic adaptive systems during run time can be found in [25], [26].

C. Adaptive Component Model

Each component in the system is represented by the *DynamicAdaptiveComponent* class. Each component may provide services to other components or use services, provided by other components. The services a component provides are represented by the *ProvidedService* class. The services a component requires are specified by the *RequiredServiceReferenceSet* class, where each instance represents a set of required services for exactly one domain interface. The *ComponentConfiguration* class of the component model represents a mapping between services required and provided. If all the required services of a component configuration are available, the provided services of that component configuration can in turn be provided to other components. In the following subsections the individual parts of the component model are introduced in more detail. Afterwards, the interplay of these parts during the configuration process will be explained.

1) Dynamic Adaptive components

Each component instance within the system is represented by an instance of the class *DynamicAdaptiveComponent*, see Figure 2.

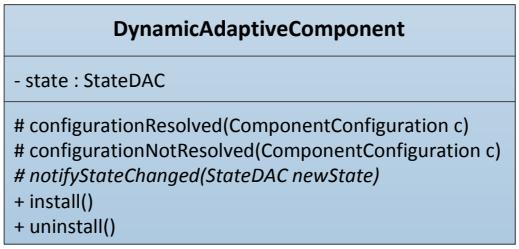


Figure 2. DynamicAdaptiveComponent class.

By calling the install or uninstall methods, a component is, respectively, published or removed from the system. If install is called, all other parts of that component are informed by calling the trigger install. The framework then starts trying to resolve dependencies on other components in order to run ProvidedServices and provide them to other components within the system. Each *DynamicAdaptive-Component* realizes a state machine, as shown in Figure 3 whose current state is stored in a variable called state.

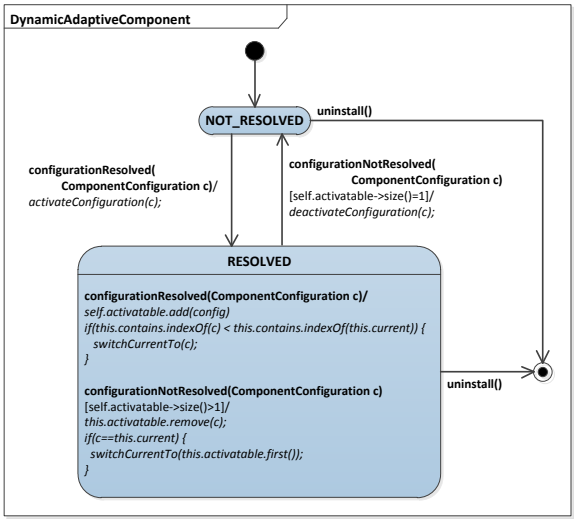


Figure 3. State machine - DynamicAdaptiveComponent class.

Two states are distinguished for *DynamicAdaptive-Component*, namely RESOLVED and NOT_RESOLVED. In the beginning a component is in the NOT_RESOLVED state. If, for a single *ComponentConfiguration*, all dependencies to services of other components are resolved, the trigger *configurationResolved* of *DynamicAdaptive-Component* is called and the state machine switches to state RESOLVED. Every time a state transition takes place, the abstract method, *notifyStateChanged*, is called. A component developer can override this method in order to react to certain state transitions, e.g., by showing or fading out a graphical user interface.

2) Component Configuration

Each component defines at least one Component-Configuration.

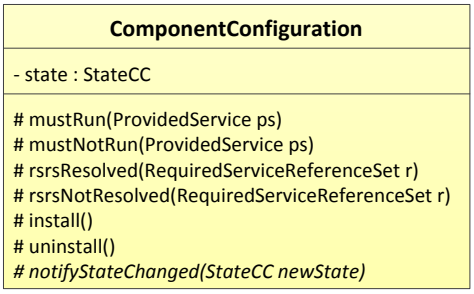


Figure 4. ComponentConfiguration class.

Figure 4 shows the corresponding class diagram for *ComponentConfiguration*. The defined *ComponentConfigu-rations* are connected to a component by the association *contains*. Each *ComponentConfiguration* represents a

mapping between a set of required and provided services. If all services required by a *ComponentConfiguration* are available, the corresponding provided services can be provided to other components. That configuration is then marked as *activatable*. In case a component has more than one *ComponentConfiguration*, an order must be defined by the component developer. During run time, at most one *ComponentConfiguration* can be active. That one is then marked as current and only those provided services are executed that are connected to *ComponentConfiguration*, which is marked as current.

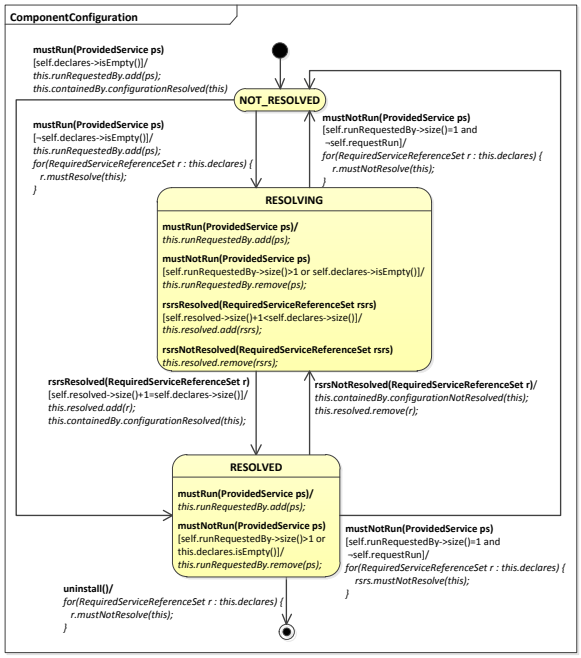


Figure 5. State machine - ComponentConfiguration class.

Each *ComponentConfiguration* realizes a state machine, as shown in Figure 5, with three states, namely NOT_RESOLVED, RESOLVING and RESOLVED. If a *ProvidedService* has to be executed (e.g., because another component needs it), the trigger mustRun of *ComponentConfiguration* is called. Afterwards the trigger *mustResolve* is called at each *RequiredServiceReferenceSet* in order to initiate the resolving of dependencies to other components. A *RequiredServiceReferenceSet* informs the *ComponentConfiguration* of the current status of the dependency resolution by calling the triggers *rsrsResolved* and *rsrsNotResolved*.

A *ComponentConfiguration* is in RESOLVED state if the dependencies of all required services are resolved, i.e., all connected *RequiredServiceReferenceSets* have called the trigger *rsrsResolved*. The *ComponentConfiguration* in turn calls *configurationResolved* to inform the *DynamicAdaptive-Component*.

3) Provided Service

A component's provided services are represented by the class *ProvidedService* shown in the class diagram in Figure 6. Each one implements exactly one domain interface. For each *ProvidedService* the number of service users who are allowed to use the service in parallel can be specified. This is done by setting the variable *maxNoOfUsers* to the required value. In our component model, a service is executed for only two reasons. The first reason is that there exist one or more components that want to use that service. Requests for service usage can be placed by calling the method *wantsUse*, or *wantsNotUse* if the usage request has become invalid. If there is a usage request for a *ProvidedService*, the connected *ComponentConfigurations* are informed by calling the trigger *mustRun*. The second reason that a service might have to be executed is that it provides some kind of direct benefit for end users. A component developer can set the flag *requestRun* in this case (e.g., because the service realizes a graphical user interface).

A *ProvidedService* realizes a state machine with three states namely NOT_RUNNING, RUNNABLE and RUNNING, as illustrated in Figure 7. A service is in RUNNABLE state if it is exclusively connected to *ComponentConfigurations* whose dependendies are resolved but none of them is marked as current. This is the case for a *ComponentConfiguration* that has higher priority and that is marked as *activatable*. However, a service is in RUNNING state if it is connected to a *ComponentConfiguration*, which is marked as *current*. If a *ComponentConfiguration* becomes current, all connected *ProvidedServices* are informed by calling the *serviceRunnable* trigger.

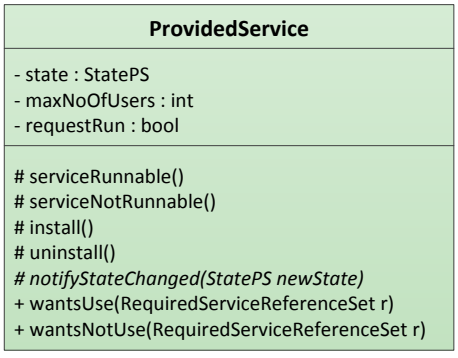


Figure 6. ProvidedService class.

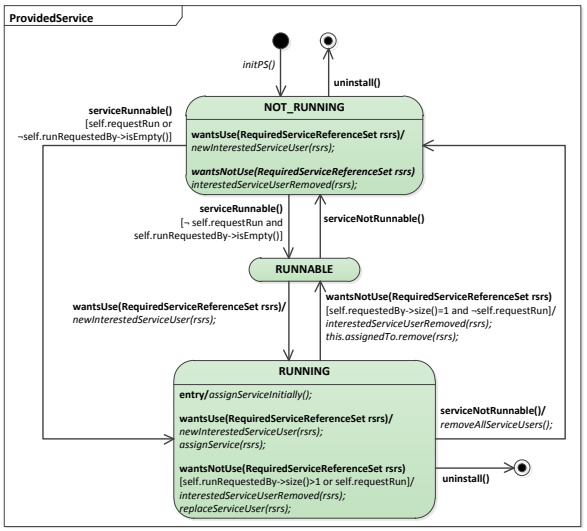


Figure 7. State machine - ProvidedService class.

4) Required Service Reference Set

A component may need functionality provided by other components in the system. In our component model those dependencies are specified with the *RequiredService-ReferenceSet* class, shown in Figure 8. Each instance of *RequiredServiceReferenceSet* represents dependencies on a set of services that implement the same domain interface. That domain interface is specified by the association *refersTo*. A component representing a trainer for example, may define a *RequiredServiceReferenceSet* that refers to a domain interface called *IAthlete* in order to get access to the training data of athletes. The minimum and maximum number of required references to services can be specified by setting the variables *minNoOfRequiredRefs* and *maxNoOf-RequiredRefs*.

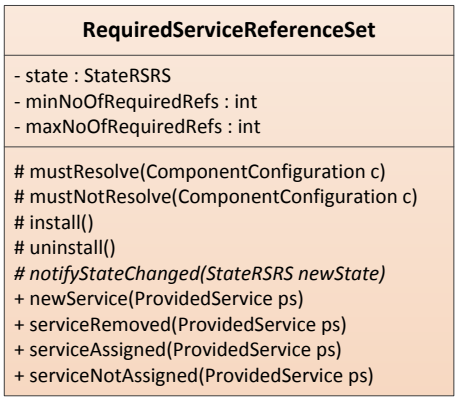


Figure 8. RequiredServiceReferenceSet class.

A *RequiredServiceReferenceSet* realizes a state machine with three states, namely NOT_RESOLVED, RESOLVING and RESOLVED. Figure 9 visualizes this state machine. As soon as there is a request for resolving dependencies, the state switches to RESOLVED or RESOLVING, depending on the value of *minNoOfRequiredRefs*. If it is zero, then the requirements are fulfilled and it can switch directly to RESOLVED. A request for dependency resolution is placed by calling the *mustResolve* trigger.

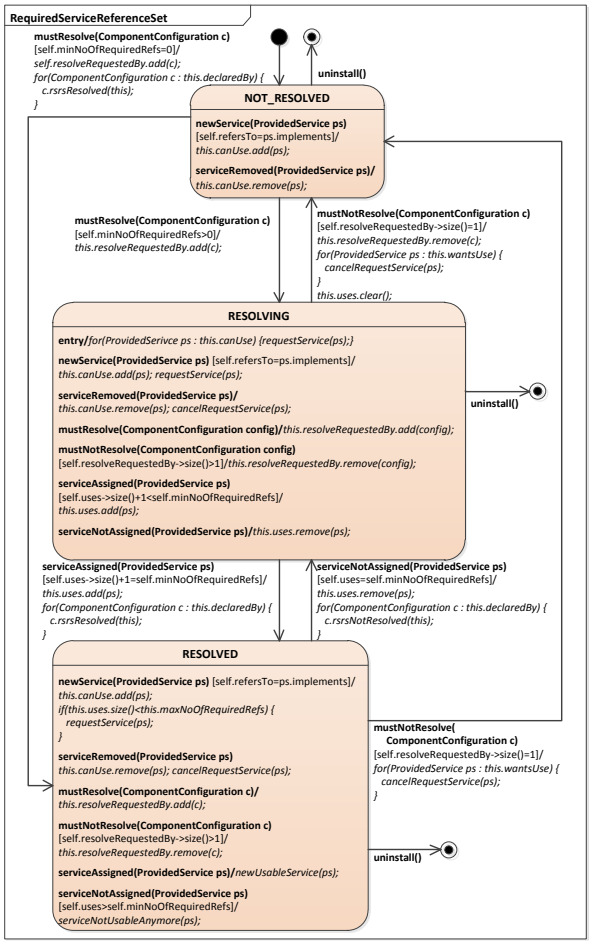


Figure 9. State machine - RequiredServiceReferenceSet class.

5) Notation for DAiSI Components

To describe DAiSI components we use a compact notation, illustrated in Figure 10. Provided services are notated as circles, required services as semicircles, component configurations are depicted as crossbars, and the component itself is represented by a rectangle. Provided services that are intended to be activated (flag *requestRun* is true) are shown as a black circle.

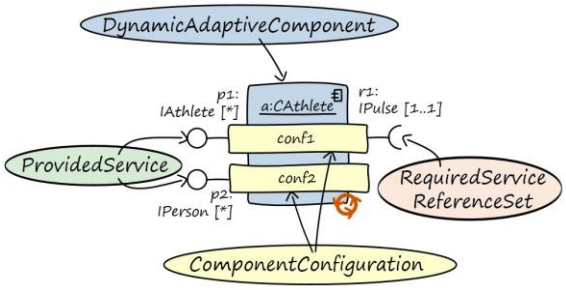


Figure 10. Notation for DAiSI components.

The component depicted in Figure 10 thus specifies two component configurations. The first requires exactly one service, which implements the *DomainInterface IPulse*. If such a service is available, the service variable p_1 of type *IAthlete* can in turn be provided to other components in the system. If no pulse service is available, the second configuration can still be activated because that one defines no dependencies to other services. In that case, the athlete component provides the service variable p_2 to other components.

D. Decentralized Dynamic Configuration Mechanism

There exist three types of relations between *RequiredServiceReferenceSets* and *ProvidedServices*, represented by the associations *canUse*, *wantsUse* and *uses*. The set of services that implement the domain interface referred by the *RequiredServiceReferenceSet* is represented by *canUse*. Note, this only guarantees a syntactically correct binding. In [25] and [26], we have shown how this approach can be extended to guarantee functional-behaviorally correct binding as well during run time using a run-time testing approach.

The *wantsUse* set holds references to those services for which a usage request has been placed by calling *wantsUse*. And the *uses* set contains references to those services, which are currently in use by the component or by *RequiredServiceReferenceSet*.

Each time a new service becomes available in the system, the newService method is called with a reference to the service as parameter. The new service is added to all *canUse* sets, if the corresponding *RequiredServiceReferenceSet* refers to the same *DomainInterface* as the *ProvidedServices*. If there is a request for dependency resolution (by a call of the *mustResolve* trigger), usage requests are placed at the services in *canUse* by calling *wantsUse* and those service references are copied to the *wantsUse* set. *ProvidedServices*

The management of these three associations—*canUse*, *wantsUse* and *uses*—between *RequiredServiceReferenceSets* and *ProvidedServices* is handled by DAiSI's decentralized dynamic configuration mechanism. This configuration mechanism relays on the state machines, presented in the previous sections, of the corresponding classes in the DAiSI framework and their interaction. In the following section, we will first describe the local configuration mechanism component and then the interaction between two components for inter-component configuration.

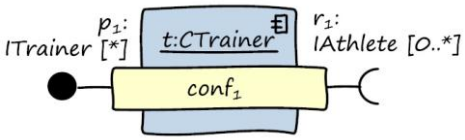


Figure 11. CTrainer component.

1) Local Configuration Mechanism

Assume a given component as shown in Figure 11. The component t of type *CTrainer* has a single configuration. It provides a service of type *ITrainer* to the environment, which can be used by an arbitrary number of other components. The component requires zero to any number of references to services of type *IAthlete*.

The boolean flag *requestRun* is true for the service provided. Hence, DAiSI has to run the component and provide the service within the dynamic adaptive system to other components and to users. As the component requires zero reference to services of type *IAthlete*, DAiSI can run the component directly and thereby provides the component service to other components and users as shown in the sequence diagram in Figure 12.

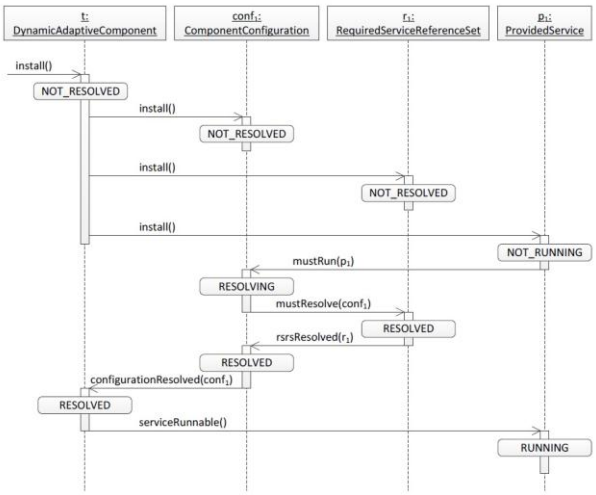


Figure 12. Local configuration mechanism component.

2) Inter-Component Configuration Mechanism

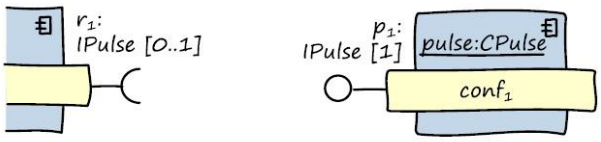


Figure 13. CAthlete and CPulse components.

Now assume two components: The *CAthlete* component, shown on the right hand side of Figure 13, requires zero or one reference to a service of type *IPulse*. The second

component, *CPulse*, shown on the left hand side of Figure 13, provides a service of type *IPulse*. Note, this service can only be exclusively used by a single component.

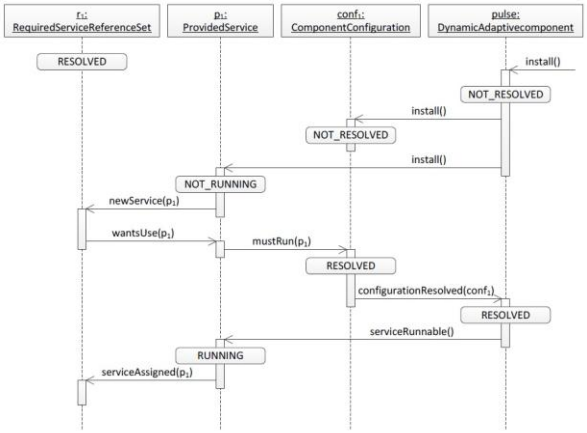


Figure 14. Inter-component configuration mechanism.

Once the *CPulse* component is installed or activated within the dynamic adaptive system, DAiSI integrates the new service in the canUse relationship of the *RequiredServiceReferenceSet* r_1 of the component *CAthlete*. Then DAiSI informs (calling the method *newService*) the *CAthlete* component that a new service that can be used is available as shown in Figure 14. DAiSI indicates that *CAthlete* wants to use this new service by adding this service in the set of services that *CAthlete* wants to use (set *wantsUse* of *CAthlete*). Once the service runs it is assigned to the *CAthlete* component, which can use the service from now on (added to the set uses of *CAthlete*).

IV. SAMPLE APPLICATION – SMART BIATHLON TRAINING SYSTEM

As already mentioned, we have realized and used a couple of dynamic adaptive systems based on DAiSI. One of the first domains for which we developed dynamic adaptive systems was training systems for athletes. For that reason we have chosen this domain to implement the first dynamic adaptive system on top of the new DAiSI version.

A. Domain Model

In the desired dynamic adaptive system, athletes (*IAthlete*) and trainers (*ITrainer*) can supervise the pulse (*IPulse*) of the athlete (see Figure 15). Moreover, athletes might use ski sticks (*IStick*), which have gyro sensors. Once connected with the sticks the athlete as well as the trainer can monitor the technically appropriate use of the sticks during skiing for the required skiing style. Once the biathlete has reached a shooting line (*IShootingLine*) he is allowed to use the shooting line only if a supervisor is available (*ISupervisor*).

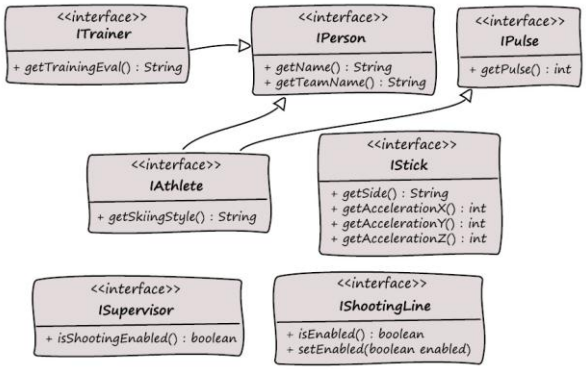


Figure 15. Domain model - "Smart Biathlon Training System".

B. Available Components

For a simple version of the system only three component types have been realized (see Figure 16): *CPulse*, *CAthlete*, and *CTrainer*. Note that additional components have been realized and evaluated for more sophisticated systems. For the purposes of this paper we only use these three components to show the decentralized configuration mechanism.

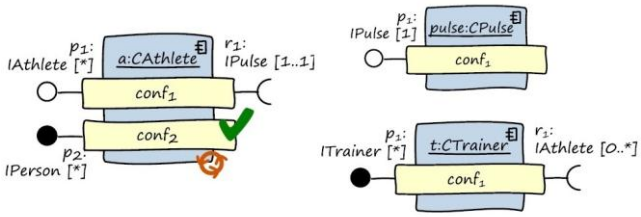


Figure 16. Adaptive components: CPulse, CAthlete, CTrainer.

The *CPulse* component provides an exclusive usable service *IPulse* and requires no other services from the dynamic adaptive system. The *CAthlete* component provides two services: *IPerson* and *IAthlete*. In $conf_2$ it provides the service, *IPerson*, which has the flag, *requestRun*, and requires no service from the environment. In conf1 it provides the service, *IAthlete*, but therefore requires a service, *IPulse*. And finally the *CTrainer* component may supervise an arbitrary number of athletes and thus provides a corresponding number of *ITrainer* interfaces to the real trainer, supporting him with the online training information of the supervised athletes.

C. Decentralized Dynamic Configuration Mechanism

Assume the following situation in the dynamic adaptive system. The component, *CPulse*, is activated and the component, *CAthlete*, is activated (see Figure 17). As the *requestRun* flag of the provided service of $conf_2$ is set and no additional service references are needed, this configuration is activated and the service is provided within the dynamic adaptive system.

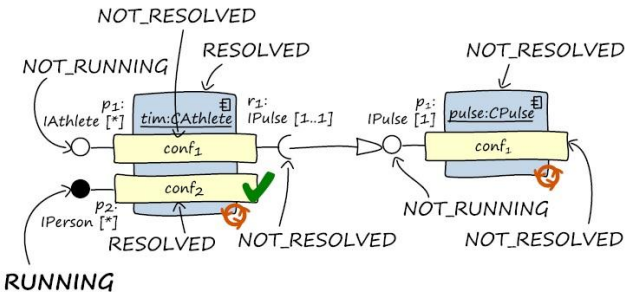


Figure 17. Initial situation in the Dynamic Adaptive System.

For the better configuration, *conf*$_1$, *CAthlete* requires a reference to a service of type *IPulse*. The *CPulse* component is able to provide this service. As the provided service, *IAthlete*, of configuration *conf*$_1$ of component *CAthlete* is not requested by any other component and has not set the *requestRun* flag, this higher configuration is not activated.

Figures 18 to 29 show the following situation: A component, *CTrainer*, has been activated and integrated into our dynamic adaptive system.

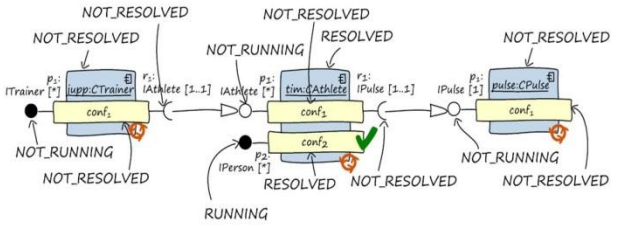


Figure 18. The CTrainer component "jupp" is deployed.

In the following the decentralized dynamic configuration mechanism is shown. Based on the interaction between the state machines of the adaptive components the dynamic adaptive system is reconfigured and the component is dynamically integrated into the system.

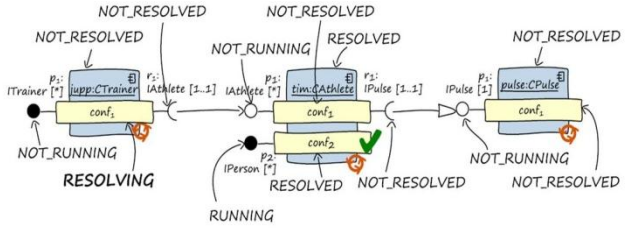


Figure 19. The component configuration conf$_1$ of Jupp's CTrainer component switches its state to RESOLVING.

The configuration strategy is then as follows. Each service with *requestRun* flag set—in Figure 18 the new service *ITrainer* of the *CTrainer* component—resolves the required services transitively from the root to the leaf.

Figures 18 to 23 show how the involved components are switched to the state RESOLVING.

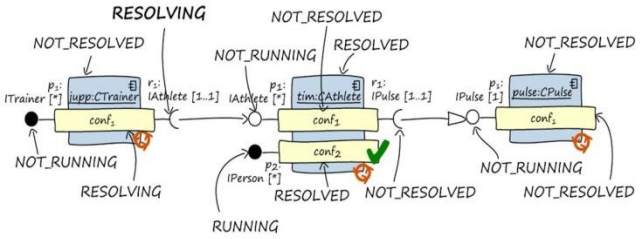


Figure 20. The interface IAthlete of Jupp's CTrainer component switches its state to RESOLVING.

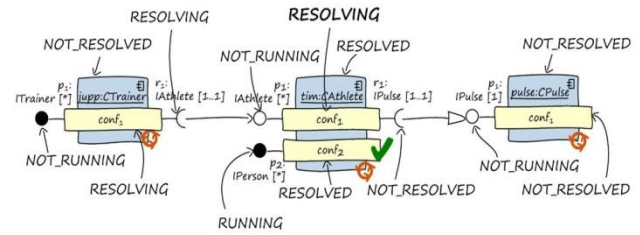


Figure 21. The component configuration conf$_1$ of Tim's CAthlete component switches its state to RESOLVING.

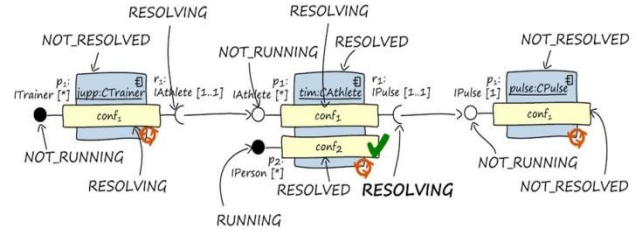


Figure 22. The interface IPulse of Tim's CAthlete component switches its state to RESOLVING.

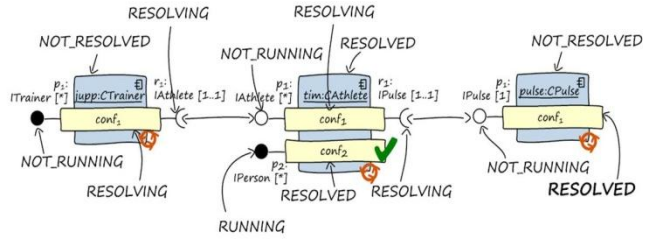


Figure 23. The component configuration conf$_1$ of the pulse component is marked as RESOLVED, because it has no required services.

Once all required services are resolved these services are activated (RUNNING) from the leaf to the root. This can be seen in Figures 24 to 29 of the application example. If not all required services were resolvable, the resolved services are set back to NOT_RESOLVED. This allows other services to resolve these services and frees the reserved resources.

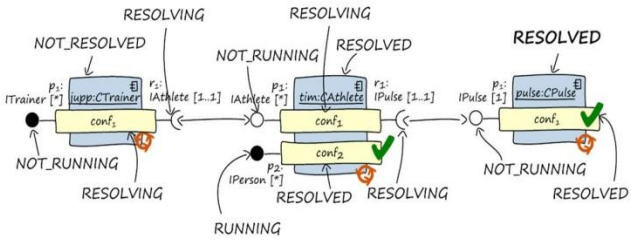


Figure 24. The pulse component is marked as RESOLVED.

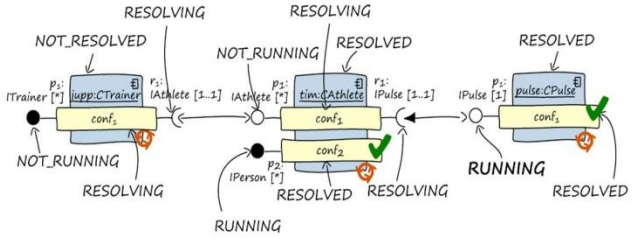


Figure 25. The IPulse interface is now RUNNING, because its requirements are resolved and a consumer (Tim's athlete component) is present.

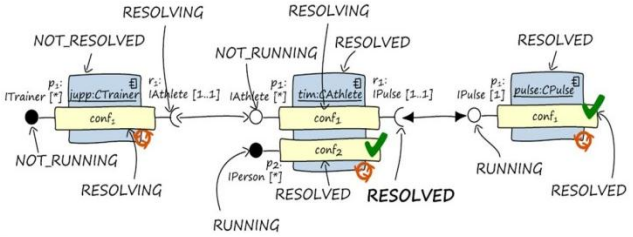


Figure 26. The required IPulse service of Tim's CAthlete component switches its state to RESOLVED.

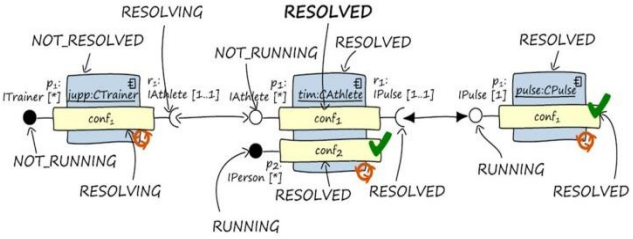


Figure 27. The component configuration $conf_1$ of Tim's CAthlete component is marked as RESOLVED.

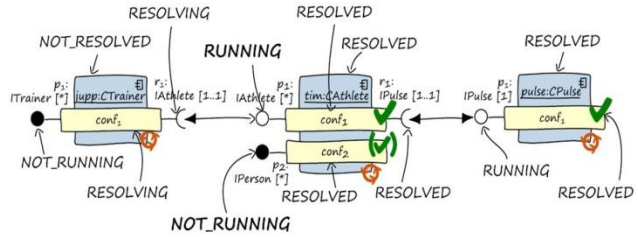


Figure 28. The provided IAthlete interface is marked as RUNNING, IPerson is now NOT_RUNNING, as the active component configuration changed from $conf_2$ to $conf_1$.

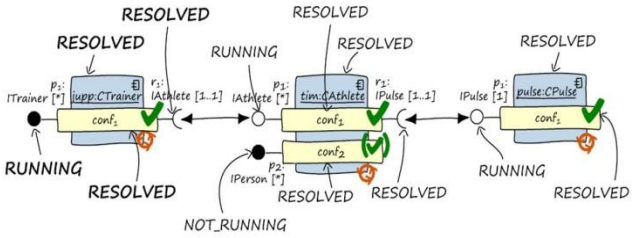


Figure 29. The configuration process is finished. Jupp's CTrainer component is now in the state RESOLVED, together with its component configurations and required services. Jupp's ITrainer interface is RUNNING.

V. CONCLUSION AND FUTURE WORK

The DAiSI approach is that a developer does not have to implement a whole dynamic adaptive system on his own. Instead the developer can develop one or more components for a specific domain. This is only possible if a domain model is available as described. This domain model has to define the interfaces between the adaptive components of the dynamic adaptive system in the specific domain.

Based on this, the developer can develop even a single component and define which interfaces from the domain architecture are required or provided in the different configurations of this component. Moreover, one can develop mock-up components providing the required interfaces in order to test the new component during development.

To support the component development DAiSI comes with two implementation frameworks. These frameworks provide several helper classes enabling a quick implementation of dynamic adaptive systems in Java as well as in C++, concentrating on the functional features of the component to be developed. DAiSI-based dynamic adaptive systems can be distributed across various machines. DAiSI is also able to establish dynamic adaptive systems across language barriers—Java- and C++-based DAiSI components can be linked together through DAiSI to form a dynamic adaptive system.

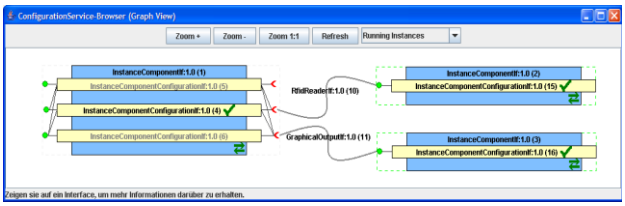


Figure 20. DAiSI Dynamic Adaptive System Monitor.

In order to monitor and debug a DAiSI-based dynamic adaptive system during development, the developer may use the so called "Dynamic Adaptive System Configuration Browser." This allows viewing the internal structure of the dynamic adaptive system in a graphical tree view.

As discussed in the introduction, DAiSI was used to realize and evaluate a couple of different applications. This allowed two main drawbacks of DAiSI to be identified: lack

of service cardinalities and the centralized configuration mechanism.

In this paper, we have shown DAiSI's new component model supporting service cardinalities and the new decentralized dynamic configuration mechanism. The decentralized configuration mechanism is needed, in order to improve performance and fault-tolerance, because of the omitted centralized configuration service. Service cardinalities are called for to increase applicability, because real-life systems often have limitations regarding the amount of service users of their provided services, and may require more than exactly one service of a given type. A first dynamic adaptive system has been successfully implemented in the assisted sports training domain.

Consequently, further systems will be realized based on the new DAiSI version. Additional research is required to establish concepts to provide a proper balance between controllability of the system's applications and the autonomy of the system components participating in these applications. To further increase applicability, more research will be put into the introduction of interface roles, to be able to make additional constraints on available provided services, used to satisfy required services.

REFERENCES

[1] H. Klus and A. Rausch, "DAiSI-a component model and decentralized configuration mechanism for dynamic adaptive systems," in Proceedings of ADAPTIVE 2014, The Sixth International Conference on Adaptive and Self-Adaptive Systems and Applications, 2014.

[2] L. Northrop, P. Feiler, R. P. Gabriel, J. Goodenough, R. Linger, T. Longstaff, R. Kazman, M. Klein, D. Schmidt, K. Sullivan, and K. Wallnau, "Ultra-large-scale systems—the software challenge of the future," Software Engineering Institute, Carnegie Mellon, Tech. Rep., June 2006.

[3] J. Kramer and J. Magee, "A rigorous architectural approach to adaptive software engineering," Journal of Computer Science and Technology, vol. 24, no. 2, pp. 183–188, 2009.

[4] C. Szyperski, "Component Software," Addison Wesley Publishing Company, 2002.

[5] M. P. Papazoglou, "Service-oriented computing: concepts, characteristics and directions," in Proceedings of the 4th International Conference on Web Information Systems Engineering (WISE 2003). 10-12 December, Rome, Italy: IEEE Computer Society Press, 2003, pp. 3–12.

[6] J. Magee, J. Kramer, and M. Sloman, "Constructing distributed systems in conic," in IEEE Transactions on Software Engineering vol. 15, no. 6, pp. 663–675, 1989.

[7] J. Kramer, "Configuration programming: a framework for the development of distributable systems," in Proceedings of IEEE International Conference on Computer Systems and Software Engineering (COMPEURO 90). 8-10 May 1990, Tel-Aviv, Israel: IEEE Computer Society Press, 1990. ISBN 0818620412, pp. 374–384.

[8] J. Kramer, J. Magee, M. Sloman, and N. Dulay, "Configuring objectbased distributed programs in rex," Software Engineering Journal, vol. 7, no. 2, pp. 139–149, 1992.

[9] R. R. Aschoff and A. Zisman, "Proactive adaptation of service composition," in: H. A. Müller, L. Baresi (Eds.): Proceedings of the 7th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS'12): Zürich, Switzerland, June 4-5, 2012. Los Alamitos, California: IEEE Computer Society Press, 2012, pp. 1–10.

[10] A. Rasche and A. Polze, "Configuration and dynamic reconfiguration of component-based applications with microsoft .NET," in Proceedings of the 6th IEEE International Symposium on Object-Oriented Real-Time Distributed Computing (ISORC 2003). 14-16 May 2003, Hakodate, Hokkaido, Japan: IEEE Computer Society Press, 2003. ISBN 0-7695-1928-8, pp. 164–171.

[11] M. Anastasopoulos, H. Klus, J. Koch, D. Niebuhr, and E. Werkman, "DoAmI—a middleware platform facilitating (re-) configuration in ubiquitous systems," in Proceedings of the Workshop on System Support for Ubiquitous Computing (UbiSys), 2006.

[12] H. Klus, D. Niebuhr, and A. Rausch, "A component model for dynamic adaptive systems," in Proceedings of the International Workshop on Engineering of software services for pervasive environments (ESSPE 2007), 2007.

[13] D. Niebuhr, H. Klus, M. Anastasopoulos, J. Koch, O. Weiß, and A. Rausch, "DAiSI—dynamic adaptive system infrastructure," Technical Report Fraunhofer IESE, 2007.

[14] H. Klus, D. Niebuhr, and A. Rausch, "Dependable and usage-aware service binding," in Proceedings of the third International Conference on Adaptive and Self-Adaptive Systems and Applications (ADAPTIVE 2011), 2011.

[15] D. Niebuhr, C. Peper, and A. Rausch, "Towards a development approach for dynamic-integrative systems," in Proceedings of the Workshop for Building Software for Pervasive Computing. 19th Conference on Object-Oriented Programming Systems, Languages and Applications (OOPSLA), 2004.

[16] "Intelligent beer dispensing system," Webpage of the cebit exhibit 2010. [Online]. Available from: http://www2.in.tu-clausthal.de/~smartschank/systembeschreibung.php, accessed 2014.12.01.

[17] „DIRMEIER SmartSchank, intelligent beer dispensing system," DIRMEIER GmbH. [Online]. Available from: http://www.dirmeier.de/DIRMEIER-0-0-0-1-1-1.htm, accessed 2014.12.01.

[18] D. Herrling, "Deriving a framework from a number of dynamic adaptive system infrastructures," Master Thesis, TU Clausthal, Clausthal-Zellerfeld, 2014.

[19] C. Bartelt, T. Fischer, D. Niebuhr, A. Rausch, F. Seidl, and M. Trapp, "Dynamic integration of heterogeneous mobile devices," in Proceedings of the Workshop in Design an Evolution of Autonomic Application Software (DEAS 2005), ICSE 2005, St. Louis, Missouri, USA, 2005.

[20] T. Jaitner, M. Trapp, D. Niebuhr, and J. Koch, "Indoor simulation of team training in cycling," in ISEA 2006, E. Moritz and S. Haake, Eds. Munich, Germany: Springer, Jul. 2006, pp. 103–108.

[21] "Bilateral German-Hungarian collaboration project on ambient intelligent systems." [Online]. Available from: http://www.belami-project.hu/~micaz/belamiproject/history /part1, accessed 2014.12.01.

[22] M. Anastasopoulos, C. Bartelt, J. Koch, D. Niebuhr, and A. Rausch, "Towards a reference middleware architecture for ambient intelligent systems," in Proceedings of the Workshop for Building Software for Pervasive Computing, 20th Conference on Object-Oriented Programming Systems, Languages and Applications (OOPSLA), 2005.

[23] M. Schindler and D. Herrling, "Emergency assistance system," Webpage of the cebit exhibit 2009. [Online], available from: http://www2.in.tu-clausthal.de /~Rettungsassistenzsystem/en/, accessed 2014.12.01.

[24] A. Rausch, D. Niebuhr, M. Schindler, and D. Herrling, „Emergency management system," In Proceedings of the International Conference on Pervasive Services 2009 (ICSP 2009), 2009.

[25] D. Niebuhr and A. Rausch, "Guaranteeing correctness of component bindings in dynamic adaptive systems based on run-time testing," in Proceedings of the 4th Workshop on Services Integration in Pervasive Environments (SIPE 09) at the International Conference on Pervasive Services 2009 (ICSP 2009), 2009.

[26] D. Niebuhr, "Dependable dynamic adaptive systems: approach, model, and infrastructure," Clausthal-Zellerfeld, Technische Universität Clausthal, Department of Informatics, Dissertation, 2010.

[27] A. Rausch and D. Niebuhr, "DemSy—a scenario for an integrated demonstrator in a smart city," ECas News Journal, 2010.

[28] C. Deiters, M. Köster, S. Lange, S. Lützel, B. Mokbel, C. Mumme, and D. Niebuhr, "DemSy—a scenario for an integrated demonstrator in a smart city," NTH computer science report, 2010.

[29] S. Lange, "Projektarbeit: regelüberwachung und regelbasierte konfiguration auf basis der ruleIT-methodik: modellierung einer Fallstudie," Unpublished Work, TU Clausthal, Clausthal-Zellerfeld, 2011.

[30] F. Paternò (Ed.), "Open pervasive environments for migratory iNteractive Services – Project Final Report," 2010.

[31] D. Herrling, "Projektarbeit: realisierung von zustandserhaltung bei der migration von OSGi bundles," Unpublished Work, TU Clausthal, Clausthal-Zellerfeld, 2011, available from: URL: http://sse-world.de/index.php/download_file/view_inline/24/, accessed 2014.12.01.

[32] H. Klus, D. Niebuhr, and O. Weiss, "Integrating sensor nodes into a middleware for ambient intelligence," in S. Schäfer, T. Elrad, and J. Weber-Jahnke (Eds.): Proceedings of the Workshop on Building Software for Sensor Networks. Portland, Oregon, USA: ACM 2006. ISBN 1–59593–491–X.

[33] H. Klus, D. Niebuhr, and A. Rausch, "Towards a component model supporting proactive configuration of service-oriented systems," in ICEBE '07: Proceedings of the IEEE International Conference on e-Business Engineering. Hong Kong, China: IEEE Computer Society, 2007.

[34] C. Bartelt, B. Fischer, and A. Rausch, "Towards a decentralized middleware for composition of resource-limited components to realize distributed applications," in Proceedings of PECCS 2013, 3rd International Conference on Pervasive and Embedded Computing and Communication Systems, 2013, ISBN 978–989–8565–43–3.

[35] D. Niebuhr and A. Rausch, "Guaranteeing correctness of component bindings in dynamic adaptive systems," in Proceedings of the 35th EUROMICRO Conference on Software Engineering and Advanced Applications (SEAA), Track on Service and Component Based Software Engineering (SCBSE). 2009.

[36] D. Niebuhr, A. Rausch, C. Klein, J. Reichmann, and R. Schmid, "Achieving dependable component bindings in dynamic adaptive systems – a runtime testing approach," in Porceedings of the 3rd IEEE International Conference on Self-Adaptive and Self-Oranizing Systems (SASO 2009), 2009.

[37] A. Rausch, J. P. Müller, D. Niebuhr, S. Herold, and U. Goltz, "IT ecosystems: a new paradigm for engineering complex adaptive software systems," in 6th IEEE International Conference on Digital Ecosystems Technologies (DEST 2012), 2012, ISSN 2150-4938.

[38] H. Klus, B. Schindler, and A. Rausch, "Dynamic reconfiguration of application logic during application migration," Version: 2011, in F. Paternò (Ed.): Migratory Interactive Applications for Ubiquitous Environments, Springer-Verlag London Limited, 2011, DOI 10.1007/978–0–85729–250–6_7, ISBN 978–0–85729–249–0.

[39] H. Klus, "Anwendungsarchitektur-konforme konfiguration selbstorganisierender softwaresysteme," Clausthal-Zellerfeld, Technische Universität Clausthal, Department of Informatics, Dissertation, 2013.

"Mining Bibliographic Data" – Using Author's Publication History for a Brighter Reviewing Future within Conference Management Systems

Christian Caldera, René Berndt, Eva Eggeling
Fraunhofer Austria Research GmbH, Graz, Austria
Email: {christian.caldera, rene.berndt, eva.eggeling}
@fraunhofer.at

Martin Schröttner
Institute of Computer Graphics and Knowledge Visualization
Graz University of Technology, Graz, Austria
Email: martin.schroettner@cgv.tugraz.at

Dieter W. Fellner
Institute of Computer Graphics and Knowledge Visualization (CGV), TU Graz, Austria
GRIS, TU Darmstadt & Fraunhofer IGD, Darmstadt, Germany
Email: d.fellner@igd.fraunhofer.de

***Abstract*—Organizing and managing a conference is a cumbersome and time consuming task. Electronic conference management systems support reviewers, conference chairs and the International Programme Committee members (IPC) in managing the huge amount of submissions. These systems implement the complete workflow of scientific conferences. One of the most time consuming tasks within a conference is the assignment of IPC members to the submissions. Finding the best-suited person for reviewing a paper strongly depends on the expertise of the IPC member. There are already various approaches like "bidding" or "topic matching". However, these approaches allocate a considerable amount of resources on the IPC member side. This article introduces how the workflow of a conference looks like and what the challenges for an electronic conference management are. It will take a close look on the latest version of the Eurographics Submission and Review Management system (SRMv2). Finally, it will introduce an extension of SRMv2 called the Paper Rating and IPC Matching Tool (PRIMA), which reduces the workload for both – IPC members and chairs – to support and improve the assignment process.**

Keywords–conference management, conference tools, paper assignment, matching algorithms, TF-IDF, information retrieval.

I. INTRODUCTION

Conferences and journals play an important role in the scientific world. Both are important channels for the exchange of information between researchers. The publication list of a researcher defines his standing within the scientific community. In order to ensure quality standards for these publications, submitted work go through the so called peer-review process.

An approach of finding suitable reviewers for this process has been presented at the *International Conference on Creative Content Technologies* (CONTENT), where the foundation of this article has been discussed [1].

This process is used to maintain standards, improve performance and provide credibility [2]. Today almost every conference or journal uses an electronic conference management system in order to organize this process.

In-a-nutshell the peer review process for a conference undergoes the following steps (see Figure 1):

- **Submission Phase** In the submission phase authors need to specify a certain amount of descriptive metadata, which is required for organizing the process. Besides title and co-authors these can include abstract, keywords and conference specific categorizations (e.g., poster or tutorial tracks). Finally, they need to provide their paper and optionally additional material (e.g., videos).
- **Review Phase** When the submission deadline has passed, the submitted papers are distributed to the reviewers. The conference reviewers are usually members of the International Programme Committee and – depending on the size of the conference – a pool of external experts. Each reviewer receives a certain amount of submitted papers depending on his expertise and workload. Further he must not be in any kind related to the author to prevent a conflict of interests. Assigning the submitted papers to the IPC members is a crucial task in the peer review process because their reviews decide, if the paper is accepted or not.
- **CRC Phase** In case of acceptance the author is allowed to upload a camera-ready-copy (CRC) version of the paper, which is then published in the proceedings of the conference.
- **Rejected** If the paper is not accepted, the submissions enter a special rejected phase. Rejected submissions are not further considered in the review process and are only used for statistical purposes (e.g., total amount of submissions in the conference).

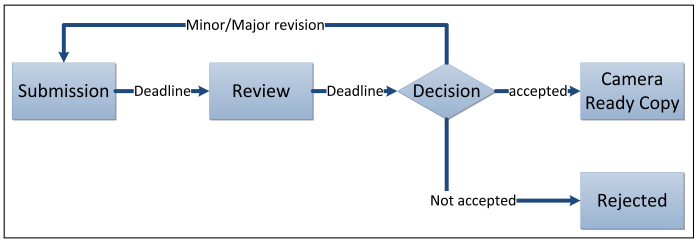


Figure 1. The typical phases of a conference and how submissions pass through this conference.

This is the essence of the peer review process. Within the peer review process there exist variations, which mostly differ in what information is revealed to whom. The most commonly used are the single and double blinded peer reviewing process:

- **Single Blind** In the single blinded peer review the identity of the reviewer is unknown to the author. But, the reviewer knows the identity of the author. In this setting, the reviewer can give a critical review without the fear that the person itself will be targeted by the author.
- **Double Blind** In the double blinded peer review, the identity of the reviewer and author is unknown to each other. This process guarantees the same chances for unknown and famous scientist and universities by not putting the name on the paper.

There are further versions of peer reviewing like open peer reviewing or additions like post-publication peer reviewing, but they are rarely been applied [3][4].

Although there are many criticisms about peer review [5], the following quote from Mayur Amin from Elsevier at the APE (Academic Publishing in Europe) Conference, Berlin in January 2011 describes the current situation:

> Peer review is not perfect, but it's the best we have.

One particular point of criticism is the poor referee selection [6]. Especially conferences with a large number of submitted papers experience an enormous time pressure for finding suitable reviewers.

How can a conference managing system support the conference chair during the reviewer assignment phase? One idea is to utilize information available from different sources about the particular persons in order to find a suitable reviewer or identify conflicts of interest. Especially bibliographic data can be a valuable source of information for finding suitable reviewers. Figure 2 shows the image section of data sources related to publications within the Linking Open Data cloud. An example of such a data source is the DBLP [7], which provides bibliographic information on computer science.

Another bibliographic service is Microsoft Academic Search [8]. Figure 3 shows the visualization of a co-author graph from Microsoft Academic Search. If a person is a direct neighbour in the co-author graph of an author, this person does most likely have a conflict of interest with the author and cannot review the paper.

II. Conference Example: Eurographics Annual Conference

Since the year 2000 the Eurographics (EG) uses the MCP system (Managing Conference Proceedings) [9] and the successor SRM (Submission and Review Management) as their conference management system. These systems have been especially tailored to support the needs of the EG.

In order to get an insight of the work of a conference chair we take a detailed look at the Eurographics Annual Conference, which is organized by the Eurographics. Figure 4 shows the number of submitted/accepted papers over the last 14 years. The requirement of at least four reviews for each paper leads to approximately more than 1000 review assignments. Assuming that the average workload of a reviewer should not exceed five reviews means that at least 200 suitable (and willing) persons have to be found.

How does the review process work in detail? The submitted papers are distributed to the members of the IPC. Each paper is assigned one primary and one secondary reviewer. These act as

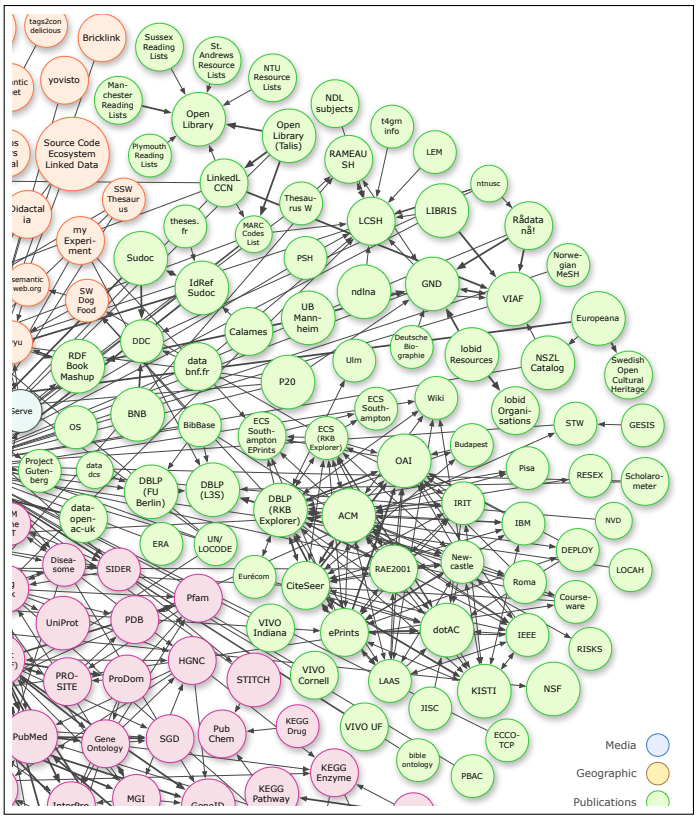


Figure 2. An excerpt of the Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. Image Source: http://lod-cloud.net

editors for this particular paper meaning they are responsible for finding at least three additional reviewers. Distributing the available submissions to the IPC members has turned out to be the most time-consuming task for the conference chairs in the last years. In order to support the distribution process the so-called "Bidding-Phase" has been introduced with SRM. IPC members are presented a list of all submitted papers (title, abstract, keywords). For each of these papers the IPC member could specify one of the following options: "want review", "could review", "not competent", "conflict of interest". Based on this classifications the system creates an automatic suggestion how to distribute the IPC members as primary/secondary reviewers. It is further possible at the start of the conference to define a list of categories. To this available categories the IPC members could specify the degree of expertise. ("expert", "passing", "not competent"). These values were matched with the author-selected categories for each paper. The weighted sum of both values indicate then the appropriateness of an IPC member for that specific paper.

Although the process of peer reviewing is unquestioned within Eurographics, over the years valuable input from the chairs in order to improve the process have been made. One of the most discussed issues was the selection of suitable reviewers. Although this weighted sum works well for the distribution, the bidding values have to be entered by each IPC manually. Going through a list of more then 200 titles and abstracts is cumbersome.

Therefore, the next version of SRM should use a new

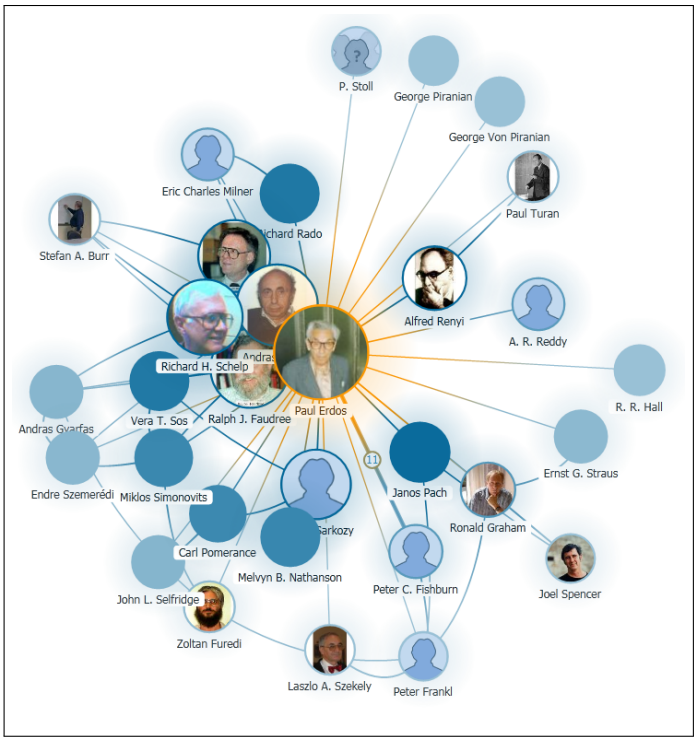


Figure 3. The co-author graph of Paul Erdös in the Microsoft Academic Search. Image Source: Screenshot in the Microsoft Academic Search http://academic.research.microsoft.com

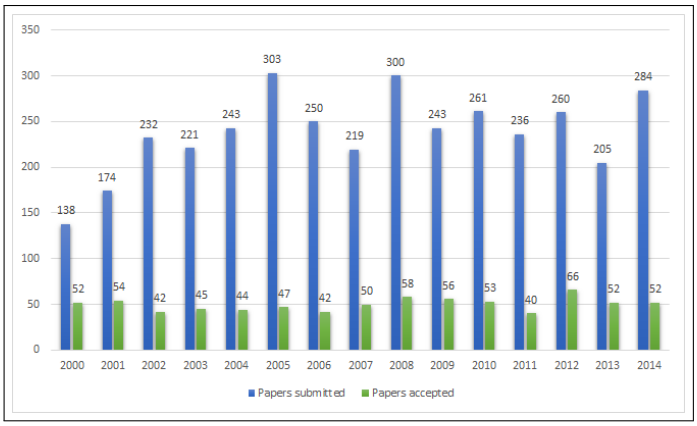


Figure 4. The submitted and accepted papers of the Eurographics Annual Conference over the last 14 years.

approach to use information available by Linked Open Data, especially bibliographic information. With the new system it should be easy to interact with 3rd party applications or data sources. It should be easy to harvest the data to create statistics and further usage of this data. Additionally, this solution should handle in a similar fashion like the current system.

III. Related Work

This section will give a small overview of some of other reviewed conference management systems:

EasyChair is a free service for managing a conference. It provides the basic features like uploading a submission and a reviewing system. It has multiple conference models a chair can choose from to customize the conference. Beside the models it is not possible to further modify the conference [10]. The review assignment process in EasyChair works manually or automatically. When using the automatic mode the International Programme Committee members define the conflicts of interests and then these members specify, which papers they are interested to review. After this is done EasyChair tries to create a good matching between the Committee members and the papers [11].

COMS Conference Management System has a one time set up fee for creating a website to satisfy the needs of the conference chair. This website will be the frontend for the chair's conference management system. Once the homepage is created the chair may define nine different review fields. The review assignment works again either manual or automatic. The automatic mode takes the reviewers biddings like in Easychair and creates a matching between the reviewers and the submissions [12].

OpenConf is a php based conference management tool, which has again the standard functionality for managing a conference. OpenConf provides the basic conference management tools. There are additional modules to add functionality to the program. One of these modules called the bidding module adds the functionality for the International Programme Committee members to define, which papers they want to review. After this bidding OpenConf provides some different algorithms to create a matching between the reviewers and the papers [13].

Confious has also the standard features for managing a conference. Confious has like the other systems an automated and a manual reviewer assignment system. But unlike the other systems Confious takes the paper topics into consideration. Authors define, which topics their paper is in and the Committee members set their experience in these topics. Then, it tries to create a good matching. Confious also tries to generate automated conflicts based on the Email and the institute of the IPC member and the author [14].

Conftool is a tool, which provides many different languages to manage a conference. Like Confious its automated review assignment takes the IPCs bidding and the paper topics into consideration when creating a review assignment. It also tries to create conflicts like Confious according to the Email, the organization and the surname of the reviewer [15].

IV. The Eurographics Conference Management System

The Eurographics Association has already a long tradition in maintaining its own conference managing system. The start of the activities date back to 1997, when Prof. Fellner was the chairman of the Eurographics annual conference, during which he experienced the amount of complex work by himself. At that time, conference support systems was just coming into existence. The first Eurographics conference management system was the *Managing Conference Proceedings* system (MCP), which has been developed by Marco Zens as a part of his PhD thesis [9]. Based on this prototype an improved version called SRM (Submission and Review Management) was developed. Since then SRM received several updates, managed 25 conferences and had over 11.000 members. Over the years valuable input from conference chairs were gathered.

These suggestions and the knowledge of the current SRM system has led to a new SRM concept, which is introduced in the next section.

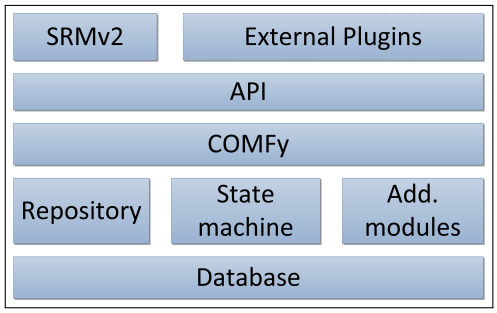


Figure 5. The concept of COMFy.

A. System Architecture

Based on the conference example in Section II the concept how the new system works is now presented. The new system provides an application programming interface (API) for managing a conference. The new SRMv2 system communicates with the core layer through a well defined API. The core layer of the new conference management framework (COMFy) provides the application logic, while the SRMv2 purely consists of the user interface. Additionally, external programs are also able to communicate with that API. COMFy itself maintains the conference data and uses the repository pattern in order to separate the business logic from the database layer. The states of a paper are represented by a state-machine. The core layer can be extended by additional modules.

The new system is divided into five different layers. The lowest layer is the database and each of the upper layers uses the functionality of the lower layers and adds additional features to the system (see Figure 5).

- **Database** The bottom layer is a relational database. Microsoft SQL Server [16] was chosen because of its filestream feature. Usually, files are stored either in the database or in a directory [17]. When stored in the database the performance to access the files is decreased drastically. On the other hand, when the file is stored in the directory the transactional consistency is lost. The filestream feature of the Microsoft SQL Server combines these features by managing the files in database management system, but stores it in a directory, which is managed by the database itself. With filestreams the transactional consistency is guaranteed and the file can be accessed fast via the directory.
- **Repositories** On the second layer there are repositories. The task of the repositories is to provide the upper layers an easy way to access the data in the database. When COMFy queries one of the repositories, this repository maps the request to an SQL statement. When the query is executed, it returns the data to the upper layer. It also works in the other direction, so the upper layer can insert new data or update existing entries.
- **State-machine** The second part of the second layer is the state machine. This is the core of the framework. It manages the phases of every submission. When a submission changes its phase it also changes the access rights of different users. For example, an author may not submit a new paper once the reviewing process starts. Another integral part of the design with the state-machine is that it is easily extensible. For example, another phase like a rebuttal phase where authors may object to the decision of the conference chair, can easily be added to the system. The current state-machine can be seen in Figure 1.
- **COMFy** This layer contains the business logic of the conference management system. It exposes these functionalities through a well defined API. It queries the repositories, parses the data and creates the response. It is also responsible for applying the different user roles, e.g., an author does not have access to the reviewer names, etc. It is designed as a model view controller pattern. So the controller takes care of the request, queries the repository, fills the model with the data from the repository and returns it to the view. Depending on the client the requested data can be delivered as Extensible Markup Language (XML), JavaScript Object Notation (JSON) or HyperText Markup Language (HTML).

B. COMFy API

The API on COMFy is based on the representational state transfer (REST) paradigm [18], which utilizes the well-known, well-defined methods of the Hypertext Transfer Protocol (HTTP) protocol (GET, POST, PUT, DELETE). This paradigm is based on clean, human readable, hierarchical uniform resource locators (URL) for accessing the resources. COMFy uses clean, structured URLs to access the requested data. The API calls can be divided in four different categories:

- **UserHome:** The “UserHome” API calls are used for retrieving information of conferences and submissions, which are tied to the user. This way the user can quickly access his own submissions or conferences.
- **Account:** The “Account” API calls are used for managing user accounts, e.g., logging into the system, registering or changing the profile information.
- **Conference:** The “Conference” API calls are for managing and viewing a conference. These calls are primarily by the chair when setting up the conference.
- **Submission:** The “Submission” calls are used for managing and viewing a particular submission. They need the conference identifier because the paper identifier is only unique within one conference. This way it is easy to identify in the URL the conference a submission is in. The calls are mainly used by authors and reviewers.

The COMFy API encapsulates the core elements of a conference system. However, an API cannot provide a clear use-case model what API calls a user needs to do for a certain task. Therefore, SRMv2 is implemented on top of COMFy as one sample application. Next to SRMv2 it is also possible to use external plugins to extend the functionality of COMFy. The *Paper Rating And IPC Matching Tool* (PRIMA) in Section V is an example, how to use this API and add extra functionality to the system.

C. Conference Setup

With the old SRM system it became clear that different chairs require different conference settings. One of the design goals of COMFy is that the conference setup should dynamic and adapt to the needs of the conference chair. The best way to show the dynamic nature of COMFy are its dynamic fields. In the two most important phases of the conference, the submission and review phase (see Figure 1) the fields can be dynamically adjusted to the conference. The following section will show how these dynamical fields work.

1) Custom Submission Fields: Figure 6 shows the submission page, where the authors enter all required meta-data for their submission. While title and abstract are common fields for all conferences, the other form fields correspond to the configured custom fields as shown in Figure 7. It is also optionally possible to upload a representative image, which identifies the submission.

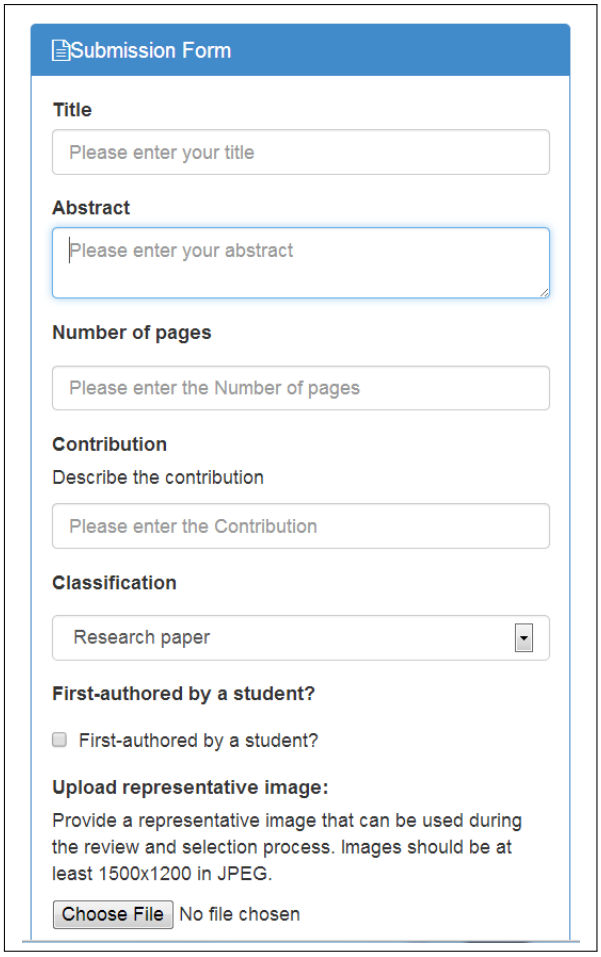


Figure 6. A dynamic generated submission page. The fields in the submission page are defined in Figure 7.

After the form has been filled out, the authors can upload their paper in the portable document format (PDF) along with additional multimedia material. Authors can modify their submission data, until the submission deadline has passed. Only the last uploaded version will be go into the reviewing process. The corresponding author receives an email for each successful upload of his paper.

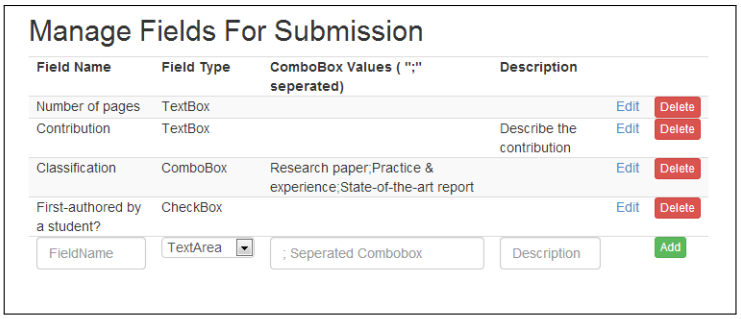


Figure 7. A possible setup for a conference. These fields have to be entered by the author when he submits a paper to the conference.

2) Custom Review Fields: Experiences from the past conferences have shown the need for customizable review forms. In order to address these requirements, SRMv2 supports four different types of custom review fields:

- **TextArea** defines a simple free-text field for the reviewer.
- **ComboBox** allows the reviewer to select a value from a predefined vocabulary.
- **ScoreBox** works like the ComboBox. But the text is matched against an integer value. This can be used, for example, like school grades to calculate the average score of the reviews and their deviation. The first value of the scorebox can be defined with the start value. Every following entry matches to the incremental integer value.
- **CheckBox** A simple checkbox, which can be ticked.

For all four types it is possible to add an optional description to the field and define the order how they appear to the reviewer. It is also possible to add a comment field for each field. This way the reviewer is able to state the reason behind his review entry. Each review has a dynamic overall recommendation and dynamic evaluation confidence. These are fixed fields, as they appear in every review. It is also possible to define, which dynamic fields should appear behind and before these two fields.

An example of a possible review setup can be seen in Figure 8. This setup generates a form for the reviewer, which can be seen in Figure 9. This review form can be downloaded as XML or HTML form. The reviewer can fill out the form offline and upload it later. In the end of the reviewing process the primary reviewer can use the overview (see Figure 10) to give the chair a senior recommendation. Or the chair can check himself/herself at the state of the reviews.

D. Review Assignment

One major challenge in the Eurographics conference is the assignment of reviewers. Within SRMv2 two different approaches exist to accomplish this task. The automated and the manual assignment. With the manual approach the chairs assign reviewers to submissions from an user-pool. In the automatic approach the program tries to create a distribution between the reviewers and the submissions.

The manual assignment can be seen in Figure 11. After choosing the submission the person who will be assigned to

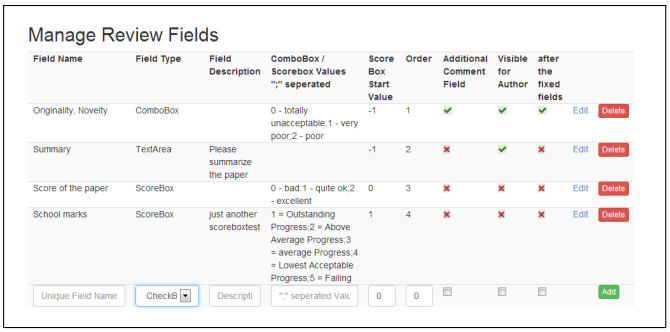


Figure 8. A possible setup for a conference. These fields have to be entered by the reviewer in order to complete his review.

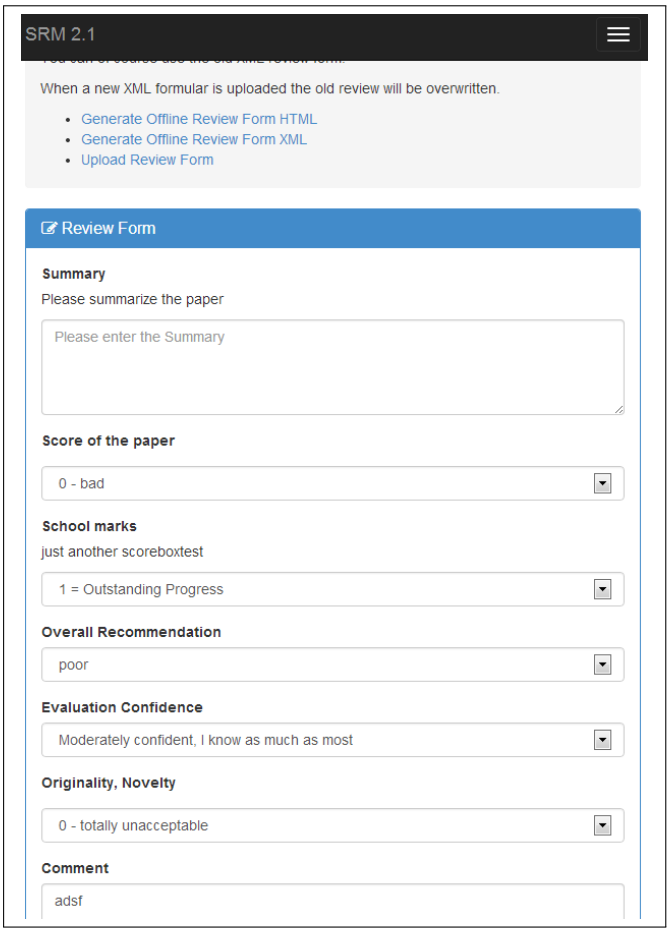


Figure 9. The example review fields setup in Figure 8 generates this review form. This review form has to be filled up by the reviewer to complete his review.

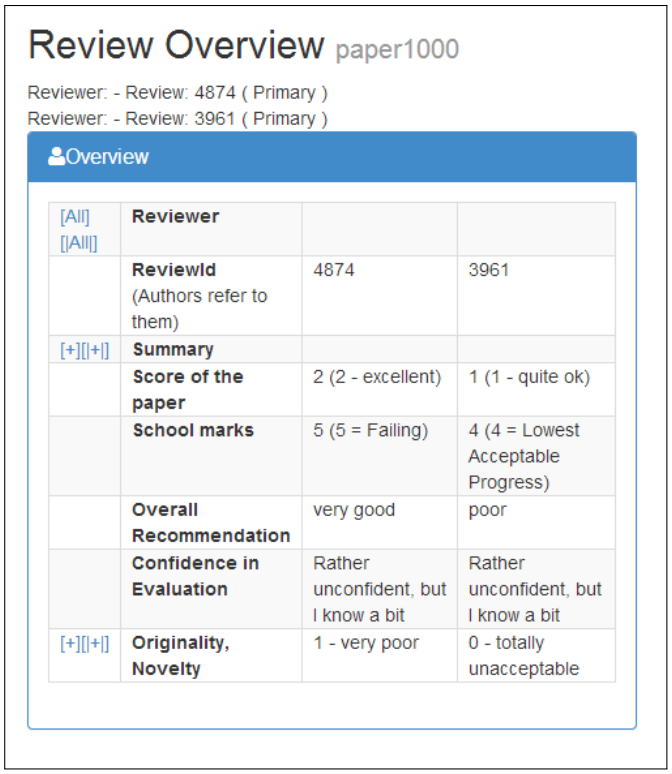


Figure 10. The review overview allows comparing the various reviews for one paper.

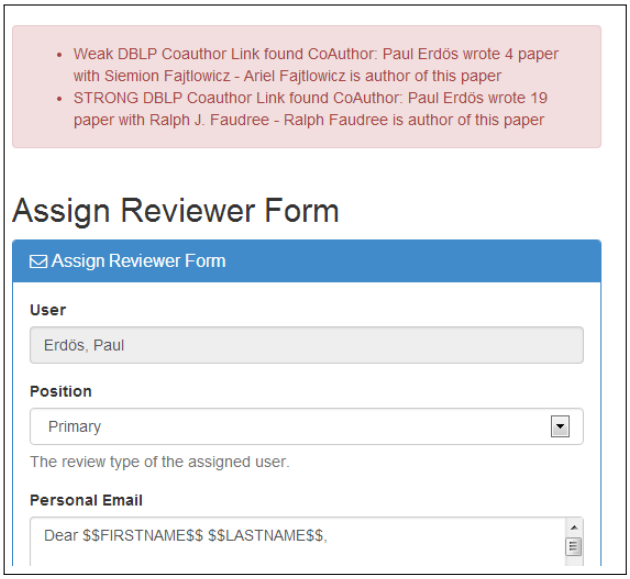


Figure 11. A warning when assigning Paul Erdös to a paper where there might be a conflict of interest with the authors of the paper.

review the paper, SRMv2 checks Linked Open Data sources like the Digital Bibliography & Library Project (DBLP) if there are conflicts of interest between the reviewer and the authors. A strong link is found, when the full names of the author and the reviewer who was selected appear in the co-author list. A weak link is found, when the domain name of the e-mail, the organization or the surname of the author and the assigned reviewer matches.

The assigning person can ignore the warning if he knows that there is no conflict of interest between the assigned person and the author. Then the assigning person has to select the user and set his reviewing role. Currently, there are three different roles: primary, secondary and tertiaries. After this task it is possible to modify the standard e-mail to create a more personalized e-mail. At last the assigning person has to confirm the assignment, so the email will be sent and the person gets

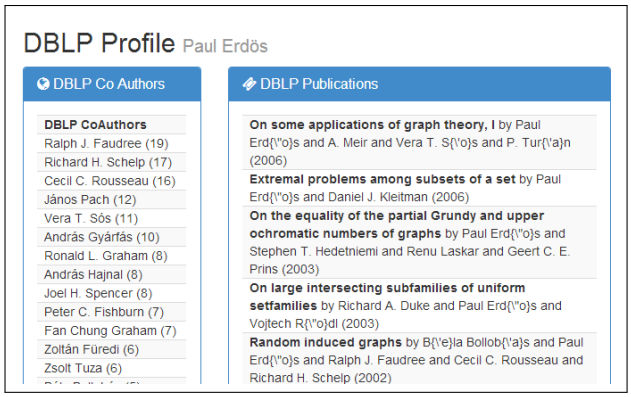


Figure 12. Harvested co-authors entries retrieved from the DBLP.

his assignment, which he can accept or decline.

A new feature implemented in SRMv2 is that the chair can now access information from the DBLP to indicate, if the person might have a conflict of interest because of a co-author relationship. In the current version the bibliographic data from DBLP is used to help identifying these conflicts. The DBLP provides an API. On this API users can query for authors. Every author in the DBLP system has an unique author identifier. After querying an author for the author identifier, it can be used to get the co-authors of that particular person. They also provide information about the amount of publications the two authors wrote together. It is also possible with the DBLP API to receive bibtex files of papers. These papers can also be found with the mentioned author pointer. Figure 12 shows the information within SRMv2 collected from the DBLP. Access to other sources like Mendeley [19] or Microsoft Academic Search are already under development.

While the first version of the Eurographics conference management system Manage Conference Proceedings (MCP) completely relied on the expertise and experience of the conference chairs, the successor SRM implemented two new features: Reviewing preferences and Bidding. With the release of SRMv2 the Paper Rating and IPC Matching Tool (PRIMA-Tool) was introduced. These approaches are described in detail in the following paragraphs.

For the automatic assignment it is necessary for IPC members to complete three steps. At first they are presented with a list of all authors. On this author list they can set their conflicts of interest with them. Then they set in, which area they are experts in. In their last step they are presented with every paper. There they set, which paper they would like to review and in, which they are not knowledgeable enough to review it. Once this is done for every IPC member COMFy tries to create the best matching of reviewer to the submission.

Before such a matching is created COMFy currently cross checks the DBLP, if there are some coauthor links, which are not defined by an IPC. If some links are found the chair is notified in the suggested matching. Currently, this system is redundant as IPC members are checking their conflicts by hand. In the future this automated assignment process will be improved and the cross checks against the DBLP should replace the manual conflict settings.

1) Reviewing Preferences: One new feature of SRM was the area of expertise list (AoE). Based on this AoE list authors could select up to five topics, where their paper fits best. In the first version of SRM all conferences shared the same AoE list, which was based on the 1998 version the ACM classification scheme[1]. Especially for some workshop series this scheme was too broadly defined, so it was decided that each conference could specify their own AoE list. The (obvious) pitfall of this decision was that now for every conference the IPC members would have to newly specify their reviewing preferences. Even for the same workshop series the AoE lists did not stay stable. Figure 13 shows an example, where the user can specify whether he is an **expert**, **knowledgeable**, **passing** or has **no knowledge** about the given topics.

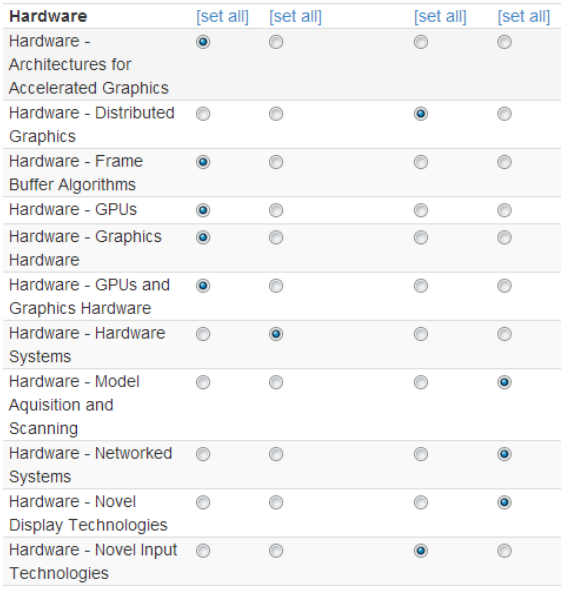


Figure 13. An IPC member can specify his knowledge in the areas of the conference.

2) Bidding: In order to support the distribution process the so-called "Bidding-Phase" has been introduced with SRM. IPC members are presented a list of all submitted papers (title, abstract, keywords). For each abstract the IPC members can specify whether they **want review (0)**, **could review (1)**, are **not competent (2)** or have a **conflict (3)**. Figure 14 shows the resulting paper/IPC matrix.

SRM uses a weighted sum combining the *Reviewing Preferences* and the *Bidding*:

$$IPCsuggestion = aoeRating + biddingRating \quad (1)$$

The first term in the sum of Equation (1) calculates the score from the IPC reviewing preferences and the categories entered by the author:

$$aoeRat. = 25 \cdot \frac{4 \cdot \#\{e\} + 2 \cdot \#\{k\} + \#\{p\}}{\#\{Area \quad of \quad expertises \quad of \quad Paper\}} \quad (2)$$

[1] http://www.acm.org/about/class/1998

Figure 14. Bidding matrix showing for each submission the bidding value entered by the IPC members. 0 means they want to review this paper, 1 that they could review the paper, 2 that they are not competent and 3 that there is a conflict of interest on one of the authors. If there is no value on the bidding matrix, then there is no information about this author on this paper.

The intersection of categories where the IPC member is expert in and the paper categories is defined as e. k is the intersection, where the IPC member is knowledgeable and p, where the IPC has only passing knowledge. The resulting *aoeRating* will be a value between 0 and 100.

The second term comes from bidding and maps the user input to a value between 0 and 100 (see Equation (3)):

$$biddingRating = \begin{cases} 100 & Want\ Review \\ 80 & Could\ Review \\ 0 & no\ Expertise \end{cases} \quad (3)$$

V. Paper Rating And IPC Matching Tool - PRIMA

Paper Rating and IPC Matching Tool (PRIMA) is the third option for IPC members to define their reviewing preferences. It is a standalone extension to SRMv2. PRIMA uses the *Term Frequency Inverse Document Frequency* [20] (TF/IDF) algorithm in order to calculate the similarity between all submitted papers and previous papers of the IPC members for extracting a matching value allowing an automatic distribution of submissions to IPC members.

Figure 15 shows the workflow of the automatic score generation with PRIMA. In the first step, the PRIMA tool is initialized with the required data for the IF/IDF calculation:

- The submitted paper along with their meta-data
- Information about the IPC members of the selected event.

PRIMA uses the API of the SRMv2 framework [21] in order to fetch the required information. After the initialization, the IPC members are invited by e-mail to upload their publications fitting the scope of the conference. The more papers a user uploads into PRIMA the better the algorithm can find different matchings to the submissions of the conference. After all initial data is available (submitted papers of the conference and the uploaded publications of the IPC members), the paper scores are calculated. These scores are then transmitted to SRMv2 in order to support the pre-ordering for the bidding process and to support the automatic assignment proposal.

Before the calculation itself starts some preprocessing steps are necessary to improve the TF/IDF result:

- **Text extraction:** For all uploaded publications, the raw text is extracted from the PDF documents. The extracted text still contains a large number of unnecessary information, which do not have an impact on the paper classification, for example, numbers, special characters, code, URLs, e-mail addresses, punctuation, authors, addresses, IDs, etc. Future work on PRIMA concentrates on further separating the text, which is useful for the TF/IDF score generation from the overhead part, which interferes with the generation [22].
- **Removal of stop words:** Stop words are words, which occur often in a text but do not add any informational value to the text. Some examples of this stop words are: *and*, *or*, *the*, *an*, *important*, *however*, *just* and so on. All these words are necessary for the creation of sentences. But two texts do not relate strongly to each other just because they have a lot of "and" together [23].
- **Stemming:** Stemming reduces words to their common root. For example, "overview" and "overviews" are not the same words in a computational matching, so the word *overviews* is reduced to *overview*. These two words will then match in the algorithm. [24].

After the preprocessing steps PRIMA starts the TF/IDF algorithm.

A. Term Frequency Inverse Document Frequency

The TF/IDF algorithm can be separated into two parts. The first part is the *Term Frequency* part. It uses the frequency of terms in a document to classify the document. The second part of the algorithm is the *Inverse Document Frequency*. It weights the terms according to the occurrence in all other documents. The more a term is used in different documents the less information it provides for classifying a document [25]. The algorithm itself is already a quite understood and researched topic in different areas like text categorization, text analysis, mining and information retrieval techniques [20].

The function $f(t,d)$ in Equation (4) counts every term t in the document d. After it has been counted every term is normalized with the logarithm.

$$tf(t,d) = log(1 + f(t,d)) \quad (4)$$

The inverse document frequency (see Equation (5)) counts the occurrences of a term across all documents in a given document corpus. This is done by taking the logarithm of the quotient between the total number of documents $|D|$ and the amount of documents d containing the term t. A term, which occurs in every document is not useful for categorizing, so it has to be penalized for being not important in the current global text corpus. Terms, which occur in fewer documents receive a higher value with this formula.

$$idf(t,D) = log\frac{|D|}{|\{d \in D : t \in d\}|} \quad (5)$$

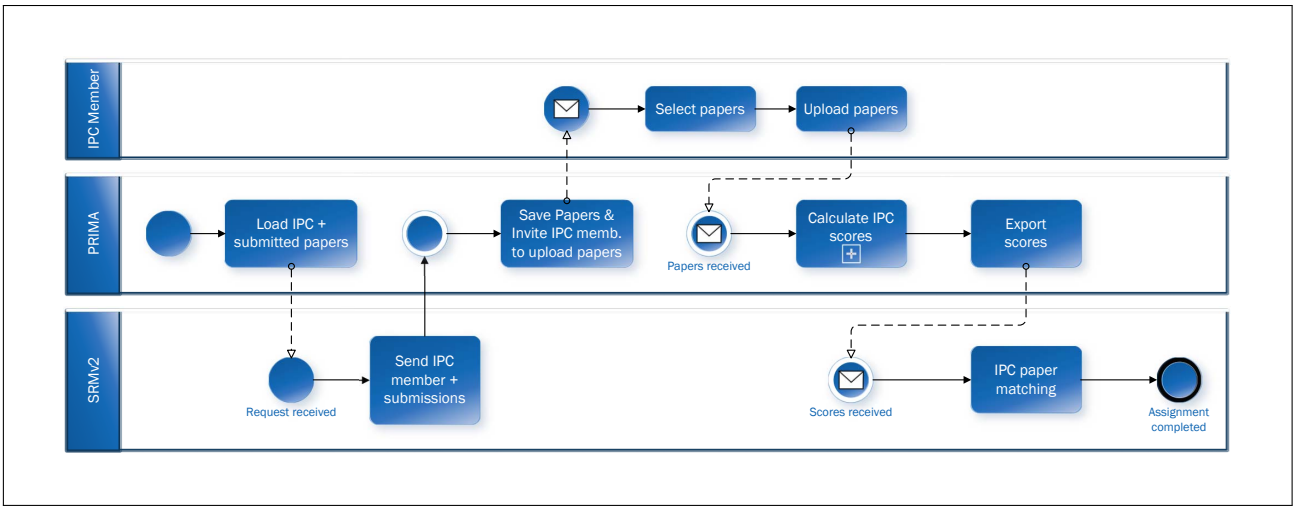


Figure 15. The complete workflow of the Paper Rating And IPC Matching Tool (PRIMA).

By multiplying the term frequency with the inverse document frequency the TF/IDF is received (see Equation (6)). This value classifies a term in a document and its classification significance across all documents [26].

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D) \tag{6}$$

All TF/IDF values of a document form a vector, which classifies the document. By calculating the angle (see Equation (7)) between two documents, it is possible to extract a similarity value [27].

$$\cos(a,b) = \frac{a \cdot b}{\| a \| \| b \|} \tag{7}$$

After the TF/IDF values are calculated, each submission is compared against all papers of the IPC members with the cosine similarity. If an IPC member has provided multiple publications, all of them are checked against a single submission paper. Currently, the average of the five best matching papers is saved. This is done to more robust against statistical outliers. Furthermore, not all papers are taken into consideration as a person might upload a lot of papers belonging to different areas.

VI. Performance and Results of PRIMA

PRIMA was first tested for the Eurographics 2014. The papers and the reviewers are anonymized and randomly reordered. The International Programme Committee consisted of 70 members and a total of 290 submissions were received. Every IPC member had entered their conflicts, defined areas of expertise to create a pre-filtering for the submissions and finally bidded on the paper. This final bidding matrix consists of 290 x 70 = 20300 entries. Figure 16(a) shows a small excerpt of the bidding matrix. The rows represent five reviewers (a to e), the columns represent 16 submissions (1 - 16). The cells are formatted with the following color scheme:

- **Light green** (0) The IPC member wants to review the paper.
- **Dark green** (1) The IPC member could review the paper.
- **Yellow** 2) The IPC member considered himself as not competent enough to review this paper.
- **Red** (3) The IPC member has a conflict with the authors of the submitted paper.
- **White** (-1) No data has been provided by the IPC member.

Most reviewers take the default *not competent* or did not submit any values at all. In the first prototype the default value for the bidding was *could review*, but due to requests from the majority of IPC members over several events this was changed to *not competent*. Therefore only a few papers contain information on the suitability of the IPC member for reviewing this paper [1].

About 300 randomly selected papers of these IPC members were uploaded and together with the 290 submissions analyses through the TF/IDF algorithm. Figure 16(b) shows the same excerpt for the TF/IDF algorithm. For each paper a value between 0 and 1 is calculated by PRIMA, where 0 means no word overlap in both documents and 1 means every word in both papers appear at the same amount. In order to archive a similar appearance like the initial bidding matrix, the following thresholds have been applied:

- **Light green** ($1 - 0.1$) High correlation between the paper and the uploaded papers of a IPC member
- **Dark green** ($0.1 - 0.05$) Medium correlation between the paper and the uploaded papers of a IPC member
- **Yellow** ($0.05 - 0.0$) Low correlation between the paper and the uploaded papers of a IPC member.
- **Red** The conflicts of the original bidding.

Figure 16(c) shows the transposed bidding and calculated matrix of reviewer C for easier comparison. The first row shows the values of the calculation, the second the bidding result of the reviewer. An important observation is, that the left bidding matrix consists of a large number of not entered information. Possible explanations are, that an IPC only checked

	A	B	C	D	E
1	-1	-1	0	2	-1
2	2	1	0	1	-1
3	-1	-1	0	2	2
4	-1	-1	2	2	-1
5	3	-1	3	2	3
6	3	-1	3	0	3
7	2	-1	2	1	-1
8	2	-1	2	1	-1
9	-1	-1	0	1	-1
10	-1	-1	0	2	-1
11	3	-1	3	1	3
12	-1	-1	0	2	-1
13	-1	-1	2	2	-1
14	-1	-1	2	0	-1
15	-1	-1	2	2	-1
16	-1	-1	2	2	-1

(a)

	A	B	C	D	E
1	0,12	0,03	0,07	0,06	0,04
2	0,07	0,07	0,03	0,16	0,32
3	0,02	0,02	0,05	0,02	0,04
4	0,05	0,02	0,03	0,06	0,22
5	0,28	0,02	0,05	0,06	0,05
6	0,03	0,03	0,02	0,12	0,11
7	0,1	0,1	0,02	0,11	0,07
8	0,1	0,03	0,04	0,23	0,06
9	0,01	0,05	0,01	0,03	0,04
10	0,05	0,03	0,21	0,14	0,07
11	0,13	0,02	0,08	0,03	0,03
12	0,04	0,02	0,01	0,08	0,19
13	0,02	0,02	0,02	0,03	0,06
14	0,08	0,05	0,06	0,32	0,05
15	0,03	0,02	0,01	0,03	0,12
16	0,01	0,01	0,01	0,02	0,03

(b)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
C	0,07	0,03	0,05	0,03	0,05	0,02	0,02	0,04	0,01	0,21	0,08	0,01	0,02	0,06	0,01	0,01
C	0	0	0	2	3	3	2	2	0	0	3	0	2	2	2	2

(c)

Figure 16. Figure (a) shows a small excerpt of the bidding matrix. Most reviewers set the values to the default not competent or did not submit any values at all. The values and colour scheme is the same like in Figure 14. Figure (b) shows an excerpt of the PRIMA matrix with the same color encoding like the bidding matrix and thresholds at 0.05 and 0.1. Figure (c) shows the transposed bidding and calculated matrix of reviewer C for easier comparison.

the papers in his own area of expertise or was not able to read all 290 submission abstracts, because of a lack of time.

Another important observation is that some good matchings are conflicts (e.g., cell $C11$ with the value 0.08 is a conflict in the left figure). It can be expected that a person who is an expert in an area also might have a project cooperation with other experts in this field and therefore has a conflict of interests with these persons.

Furthermore, it can be observed that most of the bidding values match the calculated values of PRIMA $C1$, $C3$, $C9$, $C10$ (Figure 16(c)). In addition, also the not competent column matches with the biddings $C4$, $C7$, $C8$, $C13$, $C16$. For test phase only publications from previous Eurographics events were uploaded as input for the IPC members and the amount of uploaded data also differs. For example, IPC member D had 18 uploaded papers, but person B only five. For this reason person D is much better classified by the TF/IDF and therefore has a in general better matching than person B.

Strong differences between the bidding and the calculated classification, e.g., for person C the cells $C2$, $C12$, and $C14$, can have multiple reasons. According to the TF/IDF the IPC member would be well suited as a reviewer, but he considered himself as not competent. This can have different reasons:

- The TF/IDF has analysed an older paper of the person, but the expertise focus of the person has changed.
- The title and abstract from the bidding might have been misleading.
- The submission was overlooked by the IPC member and this submissions stayed on the default value, which is *not competent*.

The first item will be addressed in further research in order to analyze if penalty value for older paper will improve the results. But also cases where the rating from the TF/IDF shows a low score, but the persons claimed that he *wants to review* occur, for example, in $C2$ and $C12$:

- Most likely the system does not have a current paper of the IPC member on this topic.
- The reviewer is interested in a paper and "wants to review" it, but does not have the necessary knowledge to review it.

The calculated values provided by PRIMA have a huge advantage: They provide indications whether an IPC member is a suitable reviewer for a given paper even if the IPC member provided no bidding information. As stated before a large portion of the bidding matrix is not filled up. In these cases it is possible to create a better reviewer-to-paper assignment instead of randomly distributing the submitted papers to the reviewers. For example, in submission 4 the best matches are person D and E, for submitted paper 13 person E would be a good choice.

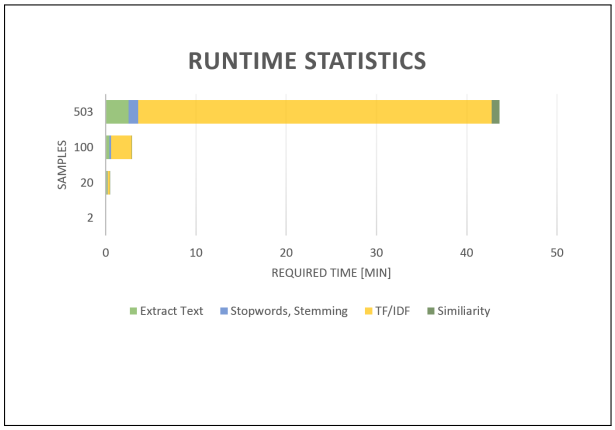


Figure 17. The amount of time each of the tasks take. It can be seen that the algorithm has an exponential growth.

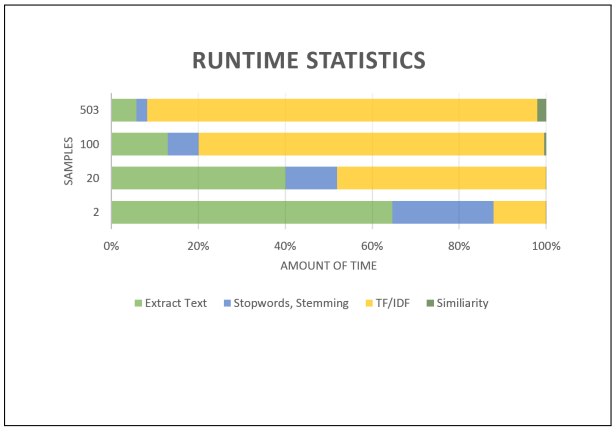


Figure 18. The amount of time each of the tasks take, split by the tasks and scaled to 100%.

Figure 17 shows the runtime statistics of the PRIMA tool split into the five steps *extract text*, *stopwords, stemming*, *TF/IDF*, and *similiarity*. It can be seen that for a small number of papers the extraction of the text and the stopwords removal and stemming takes up most of the time. If the number of papers increases, the more time the TF/IDF algorithm itself takes. The text extraction and stopwords removal and stemming can be precalculated and stored. However, this step takes less than 10% of the time during the full calculation using more than 500 papers. The TF/IDF itself cannot be precalculated as every further submission changes the weighting of each word in the calculation process. This means that the calculation can only start once all papers of all IPC members are available.

VII. Conclusion & Future Work

In this paper, we presented the COMFy conference management framework and its interaction with the PRIMA tool, which automatically calculates a ranking between submitted papers and the available reviewers. By using the TF/IDF for categorizing the submitted papers along the reviewers expertise, the workload of the reviewers and the conference chairs is reduced dramatically.

Currently, the selection and upload of publications is done manually by the reviewers. Using citation portals like DBLP [28], Citeseer [29] and other sources, the selection and retrieval of the full-text version (e.g., when available through the Open Access [30] initiative) can be automated as well.

Another important point, which might be improved is the text extraction itself. At the moment, the whole paper is used for the TF/IDF calculation. And although the numbers, special characters, URLs, stopwords, etc., are removed there are still words, which slip through, which should not be used for the analysis. For example, words like the author, the institution, figure explanations, headings, formulas and so on.

For the upcoming Eurograhics conference it is planned to use SRMv2 alongside with PRIMA and to evaluate the scores by presenting submissions to the authors in descending order. Then, the IPC member can concentrate on the title/abstracts, which fit best to the topics of his own publications. The values that the PRIMA tool generates can also be used as suggestions for the reviewer during the bidding process. This way the member can skim over the values and check if they fit. This will save the IPC members valuable time, which can be used for more important research [1].

References

[1] C. Caldera, R. Berndt, M. Schröttner, E. Eggeling, and D. Fellner, "PRIMA - towards an automatic review/paper matching score calculation," in Proceedings of The Sixth International Conference on Creative Content Technologies. IARIA, 2014, pp. 70–75.

[2] Academia Publishing, "What is Peer Review?" 2014, URL: http://academiapublishing.org/ [accessed: 2014-12-09].

[3] R. M. Blank, "The effects of double-blind versus single-blind reviewing: Experimental evidence from the american economic review," American Economic Review, vol. 81, no. 5, December 1991, pp. 1041–67, [retrieved: 03, 2014]. [Online]. Available: http://ideas.repec.org/a/aea/aecrev/v81y1991i5p1041-67.html

[4] M. W. Consulting, "Peer review in scholarly journals: perspective of the scholarly community–an international study," Author, Bristol, UK, 2008.

[5] R. Smith, "Peer review: a flawed process at the heart of science and journals," JRSM, vol. 99, 2006.

[6] D. Shatz, Peer Review: A Critical Inquiry (Issues in Academic Ethics (Paper)). Rowman & Littlefield Publishers, Inc., Nov. 2004. [Online]. Available: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/0742514358

[7] M. Ley, "Dblp: some lessons learned," Proc. VLDB Endow., vol. 2, no. 2, Aug. 2009, pp. 1493–1500. [Online]. Available: http://dl.acm.org/citation.cfm?id=1687553.1687577

[8] Microsoft Corporation, "Microsoft academic search," Dec. 2012, URL: http://academic.research.microsoft.com/ [accessed: 2014-12-09].

[9] M. Zens, "Creation, management and publication of digital documents using standard components on the internet," Ph.D. dissertation, Technische Universität Braunschweig, 2004.

[10] A. Voronkov, "Easy chair conference system," Dec. 2012, URL: http://www.easychair.org/ [accessed: 2014-12-09].

[11] N. Garg, T. Kavitha, A. Kumar, K. Mehlhorn, and J. Mestre, "Assigning papers to referees," Algorithmica, vol. 58, no. 1, 2010, pp. 119–136.

[12] M. Mandl, "Conference management system (coms)," Dec. 2012, URL: http://www.conference-service.com/ [accessed: 2014-12-09].

[13] Zakon Group, "Openconf," Dec. 2012, URL: http://www.openconf.com/ [accessed: 2014-12-09].

[14] M. Papagelis and D. Plexousakis, "Confious," Dec. 2012, URL: http://www.confious.com/ [accessed: 2014-12-09].

[15] H. Weinreich, "conftool - conference management tool," Mar. 2013, URL: http://www.conftool.net/ [accessed: 2014-12-09].

[16] G. Fritchey and S. Dam, SQL Server 2008 Query Performance Tuning Distilled, 1st ed. Berkely, CA, USA: Apress, 2009.

[17] R. Sears, C. van Ingen, and J. Gray, "To blob or not to blob: Large object storage in a database or a filesystem?" CoRR, vol. abs/cs/0701168, 2007.

[18] R. T. Fielding, "Architectural styles and the design of network-based software architectures," Ph.D. dissertation, University of California, Irvine, 2000, aAI9980887.

[19] Mendeley Ltd., "Mendeley," Jan. 2013, URL: http://www.mendeley.com/ [accessed: 2014-12-09].

[20] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval, 1st ed. Addison Wesley, May 1999.

[21] C. Caldera, R. Berndt, and D. W. Fellner, "Comfy - A Conference Management Framework," Information Services and Use, vol. 33, no. 2, 2013, pp. 119–128, [retrieved: 03, 2014]. [Online]. Available: http://dx.doi.org/10.3233/ISU-130697

[22] E. Rahm and H. H. Do, "Data cleaning: Problems and current approaches," IEEE Data Eng. Bull., vol. 23, no. 4, 2000, pp. 3–13.

[23] C. Silva and B. Ribeiro, "The importance of stop word removal on recall values in text categorization," in Neural Networks, 2003. Proceedings of the International Joint Conference on, vol. 3, 2003, pp. 1661–1666 vol.3.

[24] M. Porter, "An algorithm for suffix stripping," Program: electronic library and information systems, vol. 14, no. 3, 1980, pp. 130–137.

[25] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," Department of Computer Science, Rutgers University, 23515 BPO Way, Piscataway, NJ, 08855e, Tech. Rep., 2003.

[26] C. D. Manning, P. Raghavan, and H. Schtze. Cambridge University Press, 2008, [retrieved: 03, 2014]. [Online]. Available: http://dx.doi.org/10.1017/CBO9780511809071.007

[27] A. Huang, "Similarity Measures for Text Document Clustering," in New Zealand Computer Science Research Student Conference, J. Holland, A. Nicholas, and D. Brignoli, Eds., Apr. 2008, pp. 49–56. [Online]. Available: http://nzcsrsc08.canterbury.ac.nz/site/digital-proceedings

[28] M. Ley et al., "DBLP Computer Science Bibliography," 2013, URL: http://www.informatik.uni-trier.de/~ley/db/ [accessed: 2014-12-09].

[29] The Pennsylvania State University, "CiteSeer," 2014, URL: http://citeseerx.ist.psu.edu/ [accessed: 2014-12-09].

[30] Georg-August-Universität Göttingen Niedersächsische Staats- und Universitätsbibliothek Göttingen, "Open Access," 2013, URL: http://open-access.net/ [accessed: 2014-12-09].

A Method for Establishing Information System Design Practice

Dalibor Krleža
Global Business Services
IBM
Miramarska 23, Zagreb, Croatia
dalibor.krleza@hr.ibm.com

Krešimir Fertalj
Department of Applied Computing
Faculty of Electrical Engineering and Computing
University of Zagreb
Unska 3, Zagreb, Croatia
kresimir.fertalj@fer.hr

***Abstract*—Information system design and development practice did not evolve much in the last decade. Methodologies for design and development of information systems are still separating activities for design and development. Design is always done prior to the development, resulting in deliverables that cannot be reused in the development process. The deliverables of the design process are used by developers only as a blueprint of the information system. The Model Driven Architecture promised to change that by introducing model transformations. The whole idea introduced in the Model Driven Architecture raised the question of model quality. It is not possible to have a correct and complete transformation if models of the information system are not of high quality. It is very hard to achieve a sufficient level of model quality on a big project. The size of a project team makes it hard to control all contributions from designers, ensuring that these contributions comply with the project design practice. In this article, we deal with these issues by providing a method that allows control of the information system design process. This method provides guidance to designers in the project team by offering selection of patterns and transformations that are applicable to the current state of the information system design. The library of patterns and transformations represents previous design and development practice, containing knowledge developed during previous projects. The method proposed in this article allows selection of patterns and transformations that are suitable for the project, constraining and guiding the contributions of designers.**

Keywords-modeling; guidance; design; pattern; transformation.

I. INTRODUCTION

The practice of information design and development still has a number of issues that needs to be addressed. Current information system design and development practice is still mainly manual, and uses design mostly for documentation purposes. In order to improve the ratio of successful and unsuccessful projects, as well as to cut down the costs of the projects, more effort needs to be put into defining better and efficient design practices that can improve project team coordination and communication, as well as traceability and quality of the project deliverables.

This article is an extension of work done in [1]. Although the focus of this article is mainly on design practice, code development is tackled and mentioned as the result of the design process. The definition of a design project and the design process used in this article is given by Ralph and Wand [2] in the form of a conceptual model. The conceptual model includes a very precise definition of terms "design project", "knowledge", and "practice". Ralph and Wand argue that design is more important than code development, simply because design elements are better than code for communicating the rationale for structural and behavioral decisions. By giving potential applications of their conceptual model, Ralph and Wand set two challenges considered in this article:

1. Design knowledge management system - A system for storing and managing the design knowledge.
2. Design approach classification framework - A framework that enables classifying of design approaches. Such framework must give guidance for selection of a design approach and comparative research on different approaches to designers.

The proposed method is mainly based on the Model Driven Architecture (MDA), standardized by the Object Management Group (OMG) [3]. The MDA is an information system design approach based on models and model transformations. Using the MDA, an information system is designed (and developed) through several abstraction levels, from business oriented models to technically oriented models: Computational Independent Model (CIM), Platform Independent Model (PIM), and Platform Specific Model (PSM). The process of designing includes transformation of models between different levels of abstraction. Eventually, PSM is transformed into code. The promise of the MDA approach is to reduce the time and effort needed to code an information system, by refocusing on delivering meaningful details in the design activities.

However, the MDA is not perfect, and has its own issues. Gholami and Ramsin [4] are giving Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis of the MDA. Some of the issues recognized in this analysis are addressed by the proposed method:

1. Need for creation of custom transformations consumes a lot of time. Resolution of this issue must be in establishing reusable design practices.
2. Model quality issues. Without models of adequate quality, transformations cannot be successfully used and applied.

Methodologies for design and development of information systems are blueprints for processes [5] that allow a project team an organized way of designing and developing of an information system. Relying only on methodologies for design and development of information systems is not necessarily producing a model of high quality, because many of these methodologies do not incorporate design practices. For reference, the quality model given by Lange and Chaudron [6] is used. There are MDA specific methodologies that are addressing some of the MDA issues. Chitforoush, Yazdandoost, and Ramsin [7] give an overview of such MDA specific methodologies. Most of these methodologies were developed for specific projects, having built-in design practices that do not allow flexibility when needed. Some generic design and development methodologies, such as Rational Unified Process (RUP) [8][9], also rely on model based design.

The absence of the design practices can result with a model of poor quality, i.e., the model is untraceable, hard to transform and hard to analyze. One way to solve these problems is to establish design practices for the project. Established design practices must ensure that models are uniform and of high quality. According to the quality model [6], this means that all models are traceable, complete, and consistent, and that models correspond to the information system being designed and developed.

In this article, a method for establishing and imposing design practices is proposed. In the context of the MDA, establishing design practices means defining and imposing of patterns and transformations that need to be used during the design process of the information system. Reusing successful patterns and transformations from previous projects can help to establish design practices. The proposed method extends methodologies for design and development of information systems by utilizing existing OMG specifications to achieve guidance in the design process that addresses some of the MDA issues [4], and answers challenges set by Ralph and Wand [2]. The proposed method is an add-on to existing design and development methodologies. A certain level of compatibility between a design and development methodology and the proposed method is needed. Some of the MDA methodologies might be incompatible with the proposed method, since they already contain design practices. Tools used for designing and developing of an information system must have features that allow a project team to follow the method proposed in this article.

In Section II, a modeling space is defined. The modeling space allows combining all models of an information system together, providing relationship between them, and defining their purpose. In the same section, a relationship between pattern instances and models of different abstraction levels is given. Section II also includes the definition of a modeling library that contains design practice from previous projects. In Section III, current design practice in the context of generic methodologies is discussed, which helps understand how pattern instances are created during the course of the project. In Section IV, an overview of the pattern instance transformation is given. The pattern instance transformation is essential for the method proposed in this article. In Section V, the tracing and transformation language is defined. This language is used to bind pattern instances together, and help to establish tracing between model elements. In Section VI, an overview of the method for establishing the design practice is given. Section VII contains an example that presents how the proposed method works on a real life scenario.

II. Modeling Space

The proposed method deals with all models involved in the project. According to the MDA specification [3], "model transformation is the process of converting one model to another model". From a transformation point of view, the MDA deals with models that are directly involved in a transformation. This can involve at least one model. However, the transformation does not need to include all models involved in the project. From a design and development methodology point of view all models are somehow connected. All models that are part of the project need to be accessible by a tool that implements the proposed method. Many design and development tools use containers for keeping models and model elements [10] together. More than one container can be used for the project. Therefore, the tool itself must have the ability to keep relationships between modeling elements placed in different containers. The proposed method must deal with all modeling elements of the project, no matter how many containers are there. The conclusion is that all models and model elements must be observed as a part of one big modeling space.

A modeling space is a notation that can be used to represent the classification of all models that are part of the project. The modeling space can be drawn as a square box containing all possible models of a designed information system. The modeling space must follow the MDA philosophy, i.e., support different levels of abstraction given in the MDA specification.

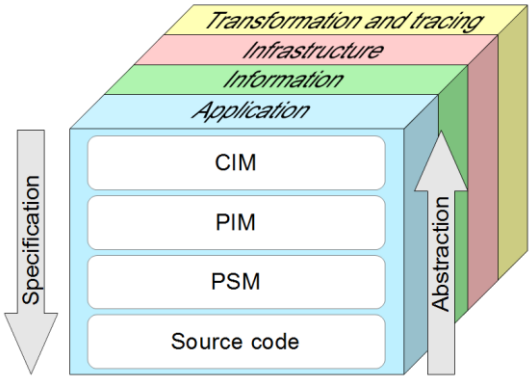


Figure 1. Structure of a modeling space

The modeling space presented in Figure 1 contains different layers, representing respective aspects or viewpoints of the designed information system. The modeling space contains four layers. The application layer is comprised of models with the business logic. The

information layer is comprised of information and data models. Models containing architecture details and infrastructure nodes are placed in the infrastructure layer. Finally, there needs to be a specific layer for transformation and tracing models. Of course, a number of layers and their purpose depend on a set of models representing an information system design. One model can belong to multiple layers. For example, a model containing requirements can easily be considered for application, information, and infrastructure related, since requirements are determining all aspects of the future information system. The modeling space must also support a clear distinction between abstract and detailed models. Abstract and computing independent models are placed on top of each layer. Models with more details are closer to the bottom of the layer.

Each model is a set of model elements. These elements originate from a modeling language, such as UML [10]. A set of models together represent the design of an information system. However, there are sets of model elements in every model that are meaningful for designers. These sets of model elements, or patterns, can be seen as reusable solutions to problems. Every level of abstraction can have its own repeating patterns of model elements. For example, CIM can contain repeating sets of model elements that can be interpreted as requirements or business processes, PIM can contain use cases or components, and PSM can contain implementation of components defined in PIM.

CIM patterns are usually created early in the project, and they depend on used architecture as well as how business analysis is performed. These high level abstract patterns have the biggest impact on the design of an information system. PIM patterns are derived from architecture and computational independent patterns. They represent an elaboration of CIM patterns within an architectural context. The most detailed are PSM patterns that represent the implementation of PIM patterns for a specific infrastructure yielded by the previously determined architecture.

A. Modeling library

In order to establish the proposed method, a library of modeling patterns and transformations must be established. The usual way to create a pattern library is by using a template document [11]. An example of online accessible pattern library can be found on [12]. This library is created by using Cloud related pattern language defined by Fehling et al. [13]. Gamma, Helm, Johnson, and Vlissides [14] propose the pattern library of basic object-oriented patterns, visualized in the UML. Hohpe and Woolf [15] propose the enterprise integration pattern library.

However, previously mentioned pattern libraries are not suitable for use by the proposed method, since solutions in these libraries are not structured, and cannot be browsed directly by a tool that implements the proposed method. In this article the Meta Object Facility (MOF) [16] family of modeling languages is used. MOF is a metalanguage standardized by the OMG. A pattern library suitable for the proposed method must utilize MOF based repository for the solution of a pattern. A good description of MOF based repository is given by Frankel [17].

The proposed modeling library can be used as the design knowledge system presented in [2]. The modeling library must have all needed features for storing and managing patterns and transformations that constitute the design knowledge.

Collecting modeling patterns can be done from existing pattern libraries, or models of already developed information systems in previous projects. Then, these patterns are inserted into the modeling library suitable for the proposed method. Collection from existing models can be done manually or automatically by detecting repetitions. Detection itself can be done by the graph matching method [18]. Pham et al. [19] propose the graph matching method for detection of cloned fragments in graph based models. According to their definition, repetitive fragments that are similar enough can be considered for clones or patterns. A similar approach can be applied to UML models.

Falkenthal, Barzen, Breitenbücher, Fehling, and Leymann [20], argue that concrete solutions are lost in the process of pattern writing. The reason for that is the need for discarding some of the solution details. If a pattern is created by using already existing information system design, then discarded model elements are the ones that need to be contributed by a designer through the process of pattern elaboration.

A pattern is a class, a blueprint that binds one or more model elements together. Application of a pattern means instantiation [20][21] within at least one model in the modeling space. Applying the pattern does not mean that the modeling is completed. Adding details and further elaboration of the pattern instance is needed, to bridge the gap between the selected pattern and final solution that was lost in the pattern writing process, which is environment and context dependent.

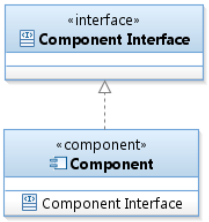


Figure 2. Example of a simple modeling pattern: a component and an interface

Figure 2 represents a pattern that is comprised of an empty interface and a component. After applying this pattern a pattern instance is created. Further elaboration of the pattern instance must add interface details, operations and parameters, subcomponents, and additional interfaces.

III. Mapping between Methodology and Pattern Sequence

Porter, Coplien, and Winn [22] have shown that pattern sequences are important, i.e., aggregation and composition of

patterns rely on the order how they are applied. A pattern library is just a set of patterns that have particular relationships among them. Even with related patterns, a slightly different order of pattern application can have different results.

The novelty introduced by the proposed method is a way to define possible pattern sequences through transformations. Transformations inside the previously defined modeling library have purpose to determine relationships between patterns, and to introduce the possibility to transform pattern instances based on patterns from the modeling library. This way, sequences are determined by transformations that are part of the modeling library.

In this article, we also argue that the selected design and development methodology (RUP for example) significantly contributes to the pattern sequence. The selected methodology defines high-level phases of a pattern sequence by providing the order how models are created and elaborated. There is a correlation between the set of transformations in the modeling library and high-level pattern sequence driven by the design and development methodology. The modeling library must contain all needed transformations that allow this high-level sequence to be completed according to the methodology.

A pattern sequence has fine course within a single project task. This fine course, or low-level sequence, is a set of activities within the task needed to complete a model, or a set of models. Transformations in the modeling library must also support these low-level sequences.

A. *Methodology driven, high-level pattern sequence*

CIMs are usually created very early in the project. In the RUP, business models are created in the Inception phase. It means that selecting and applying CIM related patterns, as well as further elaboration, can be done very early in the project. These patterns can be classified as functional requirements, non-functional requirements, business processes, or business use cases. The idea is to have these patterns and related transformations ready for use in the modeling library. Elaboration of newly created pattern instances in CIMs can be done in the Inception phase.

PIMs, part of the PSMs, architecture models, and infrastructure models, are created in the Elaboration phase. In this phase, we do most of an information system design, and take the most important decisions. In the Elaboration phase, patterns used in CIMs provide guidance for choosing patterns that could be used next. For example, usual patterns that could be used here contain use cases, components, and nodes.

The PSM is usually the last step in the design of an information system. The ultimate goal is to get the source code and deployment units. Therefore, the PSM must contain pattern instances that define a sufficient level of details for transformation into the source code, in a way that there is less work as possible for developers. Pattern instances in the PSM are mostly implementation of pattern instances in the PIM. For example, in the Component-Based Design (CBD) [23], the PSM contains platform specific implementations of components defined in the PIM.

Figure 3 provides a visual course of a project, high-level sequence of work on models and low-level sequence of pattern instantiation, transformation, and elaboration. Generally, as the project advances through the phases defined in the RUP, models become more and more specified and elaborated, until the level of actual program code. For simplicity, only one pattern instance per model is used. Models are represented by circles marked as Mi, pattern instances are represented by circles marked as Pj and transformations are represented by edges marked as t_k. Figure 3 represents an example with the following detailed low-level sequence:

1. The pattern instance P_1 is created, containing a business process. This pattern instance can be done using BPMN [24].
2. The pattern instance P_2 is created, containing a set of model elements that represents architectural decisions about selected middleware (application server, database). UML [10] can be used for this purpose.
3. Transformation t_1 is used to extract a business object from the business process information flow into the

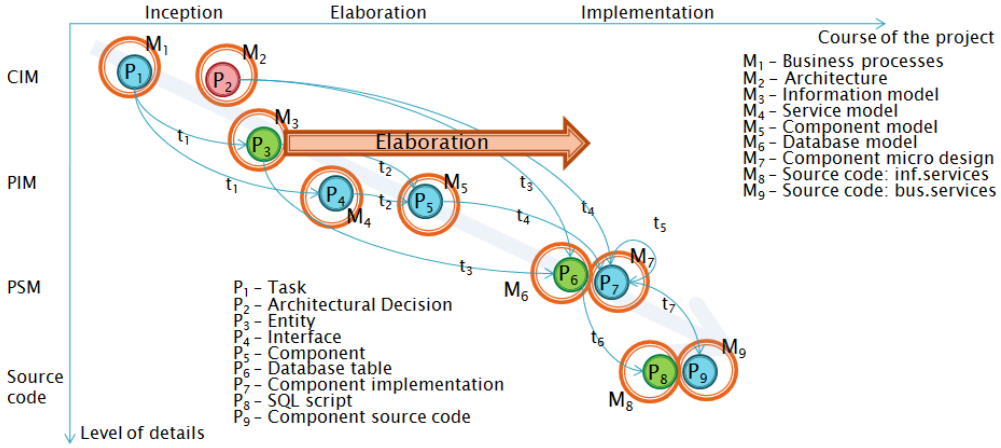


Figure 3. RUP and advancement through the design of an information system

resulting pattern instance P_3 that represents an entity of the information system. Transformation t_1 is also used to extract the interface from a task that belongs to the business process in the pattern instance P_1. As the result of transformation t_1, the pattern instance P_4 is created, representing the interface of the source task.

4. Pattern instances P_3 and P_4 are elaborated by adding operations and attributes.
5. Transformation t_2 is used to create the pattern instance P_5 from pattern instances P_3 and P_4. The pattern instance P_5 contains a component [10][23] that has an association to the entity in the pattern instance P_3, and realizes the interface in the pattern instance P_4.
6. Transformation t_3 is used to create the pattern instance P_6 from pattern instances P_2 and P_3. The pattern instance P_6 represents a table that reflects the entity in the pattern instance P_3. Transformation t_4 must take in account modeling element that contains the architectural decision about used database.
7. Transformation t_4 is used to create the pattern instance P_7 from pattern instances P_2 and P_5. The pattern instance P_7 represents implementation details of the component in the pattern instance P_5, taking into account the architectural decision about used application server.
8. Optionally, transformation t_5 can be used in case when additional standard implementation is added to the implementation of the component in the pattern instance P_7.
9. Transformation t_6 is used to create a Structured Query Language (SQL) script that can be used to create the table from the pattern instance P_6.
10. Transformation t_7 is used to create the Java code for the EJB defined in the pattern instance P_7.

As the design of an information system advances through the project, designers can create new pattern instances, or elaborate on existing ones. A new pattern instance can be created to document business need, reflect already existing functionality that can be reused, or by transforming from already existing pattern instance in the modeling space. Transformation between pattern instances is probably the most used option. Elaboration of the existing pattern instances is also very important. Once a new pattern instance has been created, it must be elaborated in subsequent project activities.

IV. PATTERN INSTANCE TRANSFORMATION

Model transformation is a key procedure in the MDA. The MDA specification [3] contains various different model-to-model transformation combinations and examples. Transformation can be done within the same model, between two different models, for model aggregation, or model separation. Grunske et al. [25] are presenting an important notion of "horizontal" and "vertical" transformations. Horizontal transformation is done between models of the same abstraction level. Typical horizontal transformation is PIM to PIM, or PSM to PSM. Any transformation within the same model is also a horizontal transformation. Vertical transformation is done between models of different abstraction levels, or from a model to the source code. A transformation from PIM to PSM, or from PSM to the source code is vertical transformation.

Model transformation can be done either manually or automatically. Manual model transformation is more common than we think. It is not unusual for a designer to start modeling from scratch by using models delivered earlier in the project. When a modeling language is structured and formal enough, automatic transformation can be used. All modeling languages derived from MOF can be transformed automatically.

Automatic transformation takes elements of a source model and converts them into elements of a target model by using transformation mapping. Transformation can be additionally used to establish relationships between models, or to check consistency of elements between source and target model. Czarnecki and Helsen [26] elaborate a number of model transformation approaches. Most used are graph based transformations and transformation languages. Transformation languages can be declarative or imperative. The OMG standardized group of MOF based transformation languages named Query/View/Transformation (QVT) [27]. QVT Relational language (QVT-R) is a typical example of a declarative approach with the graphical notation. QVT Operational language (QVT-O) is an example of an imperative approach. In this article, we are using the declarative approach.

In order to understand which transformation features are needed for the method proposed in this article, basic principles of a model transformation must be observed. Let us define a modeling space as a finite set of models

$$M_S = \{M_1, M_2, \dots, M_n\} \quad (1)$$

Each model is a finite set of elements

$$M_i = \{e_1, e_2, \dots, e_m\} \quad (2)$$

A transformation is a function

$$tr: M_S \rightarrow M_S \quad (3)$$

that takes a set of elements er_{So} from a set of source models $So \subseteq M_S$ such that $er_{So} \subseteq \bigcup So$, and translates them into another set of elements er_{Ta} in a set of target models $Ta \subseteq M_S$, such that $er_{Ta} \subseteq \bigcup Ta$. A transformation can be done within the same model $So = Ta = M_i$, or between two disjunctive sets of models $So \neq Ta$. Since a transformation can have multiple models from source and target side, these sets do not need to be disjunctive $So \cap Ta \neq \emptyset$, meaning that the transformation can include the same model M_i on source and target side, or $M_i \in So \wedge M_i \in Ta$. A transformation can use the same source and target elements, meaning that

$er_{So} \cap er_{Ta} \neq \emptyset$ when $So \cap Ta \neq \emptyset$, or it can use two disjunctive sets of elements $er_{So} \cap er_{Ta} = \emptyset$.

From a pattern point of view, each pattern instance is a set of model elements. This definition is valid for cross model pattern instances as well. All pattern instances in the modeling space M_S form a finite set of pattern instances

$$M_P = \{p_i : 0 < i \leq m \wedge p_i \subseteq \cup M_S\} \quad (4)$$

In this context, transformation is a function

$$tr : M_P \rightarrow M_P \quad (5)$$

Such transformation takes a set of source pattern instances $p_{So} \subseteq M_p$, and translates them into model elements that form a set of target pattern instances $p_{Ta} \subseteq M_p$. More precisely, transformation can be written as

$$tr : p_{So} \rightarrow p_{Ta} \quad (6)$$

Every transformation can be encapsulated in a black box implementation. Such an approach is used in [27] along with the QVT specification. According to the QVT specification, every transformation can be defined as a black box having an interface that depends on the context of transformation usage.

A. Transformation rules

Using a declarative approach for transformation of pattern instances means that every transformation can be represented as a set of transformation rules that define the relationship between a set of source model elements and a set of target model elements [26][27][28].

Czarnecki and Helsen [26] are giving important features of a declarative transformation consisting of transformation rules. As defined in [26], each transformation rule has "the left-hand side (LHS) that accesses the source model and the right-hand side (RHS) that expands in the target model". In this article, LHS is referred as "the source side" and RHS as "the target side".

Jouault and Kurtev in [29] are defining execution model for ATLAS Transformation Language (ATL) rules. Matching transformation rules in their model have a declarative part and an optional imperative part. The execution algorithm is matching the declarative part of a transformation rule, which is then fully executed (the declarative and the imperative part) in case that the transformation rule matches the supplied source pattern. It is very important to notice that declarative transformation rules are independent of each other, and that the execution algorithm does not guarantee the order of execution.

Transformation and related transformation rules, especially if they are written in a declarative way, are logic programs [30]. Transformation tr (6) can be defined as a logic program comprised of a set of rules

$$tr = \{r_1(p_{So}), r_2(p_{So}), \dots, r_n(p_{So})\} \quad (7)$$

The conditional part of every rule in previously defined set is comprised of atomic logic functions that involve model elements

$$r_i(p_{So}) \leftarrow a_1(y_1, p_{So}) \wedge a_2(y_2, p_{So}) \wedge \dots a_m(y_m, p_{So}) \quad (8)$$

where $y_j \in p_{So}$ is a model element in source pattern instances of the transformation tr. According to (8), rule r_i is matched only if all of the atoms are evaluated as true.

A transformation written in the QVT-R has two different modes: checking mode and enforcement mode. In the checking mode, transformation rules can be used to validate correctness and completeness of involved pattern instances. In the enforcement mode, transformation rules can be used for creating, updating, or deleting model elements in target pattern instances, in order to reflect all the details found in source pattern instances.

1) Applying transformation

As already defined, a transformation takes a set of modeling space model elements and translates them into another set of model elements. Earlier definition (6) shows that the transformation can include pattern instances as model element containers.

According to (8), every transformation rule consists of a set of atoms that are used to determine whether model elements in a set of source pattern instances are matching conditions of the transformation rule or not. So far, there are no additional conditions in (7) that would indicate whether a transformation can be applied to the source pattern instances or not. In order to define conditions whether a transformation can be applied or not, a set of transformations is divided on two disjunctive subsets. We define a set of "mandatory transformation rules", which need to match the source pattern instances for transformation to be applicable.

$$mtr(tr) = \{mr_1(p_{So}), mr_2(p_{So}), \dots, mr_n(p_{So})\} \subseteq tr \quad (9)$$

When a transformation is applied, is it certain that all action parts of mandatory transformation rules will be executed, if all mandatory transformation rules match supplied source pattern instances, i.e., transformation is applicable to this set of pattern instances. We also define a set of "optional transformation rules", which do not need to match the source pattern instances for transformation to be applicable.

It is possible that an atom in a transformation rule of the applied transformation matches more than one model element in the source pattern instances. A tool implementing the proposed method must allow a designer to choose which model element will be transformed. For example, a model can contain a set of use cases. The designer applies a generic transformation that can be applied to any use case. Obviously, the tool must allow him to choose which use case will be transformed by the applied transformation.

$$otr(tr) = \{or_1(p_{So}), or_2(p_{So}), \dots, or_m(p_{So})\} \subseteq tr \quad (10)$$

If conditional part of an optional transformation rule does not match supplied set of source patterns, the action part of this rule is not executed. Optional transformation rules can be used to transform elaborated details.

This might lead to a conclusion that mandatory transformation rules need to cover transformation of model elements that comprise a pattern in the pattern library, and that optional transformation rules must cover transformation of all model elements added after a pattern instance was created, i.e., elaborated model elements. While this is generally correct, mandatory transformation rules might include some elaborated details, making this transformation applicable only after elaboration of the pattern instance. This way, a designer is forced to contribute details before proceeding further in a pattern sequence.

When a transformation is applied, execution of the transformation must perform several different steps.

As the first step, the set of mandatory transformation rules of the applied transformation must be matched with the supplied source pattern instances. If all mandatory transformation rules are matched on the source side then the transformation can be applied to the supplied source, i.e., the transformation can be applied to the source pattern instances that contains all model elements needed by the mandatory transformation rules. Transformation applicability can be expressed as

$$A(mtr(tr), p_{So}) \leftarrow mr_1(p_{So}) \wedge \dots \wedge mr_n(p_{So}) \quad (11)$$

The second step is the creation of the target pattern instances. Matched transformation rules, execute their action parts creating all target pattern instances and their model elements. Every pattern is characterized by the mandatory model elements that define the essence of the pattern, or what makes this pattern different from other patterns. Redefining (8) for use in (9) results with

$$mr_i(p_{So}) \leftarrow a_{i,1}(y_{i,1}, p_{So}) \wedge \dots \wedge a_{i,m(i)}(y_{i,m(i)}, p_{So}) \quad (12)$$

where $m(i)$ is a number of atoms in i-th mandatory transformation rule. Mandatory model elements in the supplied source pattern instances for the applied transformation tr, can be expressed as

$$me(p_{So}, tr) = \bigcup_{i=1}^{|mtr\,(tr)|} \bigcup_{j=1}^{m(i)} y_{i,j} \quad (13)$$

Whether a model element in the set of target patterns created by the applied transformation is mandatory or not, can be determined only in the context of another transformation from the library. However, in the context of the transformation that created the set of target pattern instances, all model elements created by mandatory transformation rules of the applied transformation are considered for mandatory model elements.

The last step is to create a set of constraints that will disallow designers to change some of the model elements in the involved pattern instances. Transformation binds involved pattern instances together by imposing constraints on their model elements. Each pattern instance can be bound with other pattern instances through several different transformations. Constraints are imposed by the mandatory transformation rules.

Imposed constraints are used to limit designer changes in the modeling space to prevent:

1. Violating correctness and completeness of the pattern instances by changing their mandatory model elements. Obviously, all mandatory model elements must be constrained.
2. Breaking transformation bindings by changing model elements that match source and target side of the mandatory transformation rules. In this case, constrained model elements do not need to be mandatory.

One constraint can be applied to a set of model elements. Each constraint also must contain a set of forbidden actions. At the moment, it is expected that constraining updating and deleting specific model elements is sufficient.

Let us define an involved pattern instance made of l model elements

$$p_i = \{e_1, e_2, \dots, e_l\} \subseteq \bigcup M_S \quad (14)$$

and a finite set of transformations applied to p_i

$$tr_{applied}\,(p_i) = \{tr_1(\mathrm{p_i}), tr_2(\mathrm{p_i}), \dots, tr_k(\mathrm{p_i})\} \quad (15)$$

From (14) and (15), we can derive a mapping function

$$C: tr_{applied}\,(p_i) \rightarrow X \quad (16)$$

where $X \subseteq p_i$ is a set of model elements in p_i constrained by all transformations from (15). Every pair of applied transformations can constrain a different subset of model elements in p_i

$$C(tr_j(\mathrm{p_i})) \cap C(tr_k(\mathrm{p_i})) = \emptyset \wedge j \neq k \quad (17)$$

In the context of (13) and (15), a set of mandatory model elements of the pattern instance p_i can be defined as

$$me(p_i) = me(p_i, tr_{applied}\,(p_i)) = \bigcup_{j=1}^{k} me(p_i, tr_j(p_i)) \quad (18)$$

having the following condition satisfied

$$C(tr_j(p_i)) = me(p_i) \wedge C(tr_k(p_i)) \cap me(p_i) = \emptyset \quad (19)$$

The conclusion is that the set of mandatory model elements for (14) is just a subset of constrained model elements by the set of applied transformations.

$$me(p_i) \subseteq \bigcup_{j=1}^{k} C(tr_j) \quad (20)$$

Each pattern instance can be a result of several different pattern instances done earlier in the same project, or it can be a reason for creating several new pattern instances later in the same project. Several good examples can be found in [14]: a facade associated with a web service client can be used as a mediator between two different subsystems. In this example, the mediator is the pattern whose instance is bound by two different transformations.

Definition: The measure of transformation applicability

The measure of transformation applicability is a percentage of transformation's mandatory rules that match supplied source pattern instances.

We already defined transformation applicability in (11). If not all of the mandatory transformation rules are matching supplied source pattern instances, then the transformation cannot be applied. The information on how much and which rules are not matched can be very valuable for a designer. This way, the designer can see how to elaborate piece of the information system design he is working on, in order to proceed in the pattern sequence. If we define a subset of mandatory transformation rules that are matching supplied source pattern instances as

$$mtr_{matched}(tr) \subseteq mtr(tr) \quad (21)$$

then the measure of transformation applicability can be expressed as

$$MA(mtr(tr), p_{So}) = \frac{|\text{mtr matched (tr)}|}{|\text{mtr (tr)}|} \quad (22)$$

Of course, the rest of the set of mandatory transformation rules, i.e., those that are not matched

$$mtr(tr) \setminus mtr_{matched}(tr) \quad (23)$$

can be used to determine what exactly is missing in the information system design. Consulting the measure of transformation applicability is one aspect of the design guidance.

The applicability of a pattern is previously considered by Fehling et al. in [13]. The pattern library presented in their article can give the applicability of a pattern based on the context which needs to be supplied manually. In the proposed method, the context is calculated from the model space, i.e., the model space makes a transformation in the modeling library applicable or not.

2) Pattern instance elaboration

A transformation can be used to perform changes on involved pattern instances. This approach is used when new pattern instances are created, or existing instances are updated or deleted. Even when two pattern instances are bound with the transformation, the source pattern instance can be elaborated by adding new details and model elements. A transformation can be made so that these newly added details automatically update the target pattern instance.

Model elements that are not constrained by one of the binding transformations are handled by optional transformation rules responsible for spreading of elaboration details. Bidirectionality is a very important transformation aspect described in [27] and [28]. While transformation might constrain changes of some model elements in target pattern instances, changes of unconstrained model elements in pattern instances across the modeling space are encouraged. Such changes must be propagated throughout the modeling space, wherever transformation between pattern instances allows it. This propagation must be automatic and seamless.

3) Top-level pattern instances

Top-level pattern instances do not have predecessors. These pattern instances can be modeled manually by a designer without using any transformation, instantiated directly from the modeling library, or they can be created by using a transformation.

If a top-level pattern instance is instantiated directly from the modeling library, then all model elements from the selected pattern are copied from the MOF repository directly to the model in the model space. Of course, the instantiation process must rename the selected pattern model elements, and impose constraints on the newly created pattern instance. In order to know which model elements are constrained, this information must be kept together with the solution of a pattern in the MOF repository of the modeling library.

If a transformation is used, such transformation does not need to have input source pattern instances. In order to give the transformation some instructions, input parameters can be used. Transformations that create only target pattern instances can be used both for validation and enforcement purposes. All transformation rules in this transformation are mandatory transformation rules that create an initial version of target pattern instances, and impose constraints on them. Obviously, these mandatory transformation rules are always matched, even when there is no supplied set of source pattern instances. However, imposed constraints must allow elaboration of newly created top-level pattern instances in order to allow adding needed details. Functional and non-functional requirements are typical examples of top-level patterns. An external service definition is another example of such pattern.

V. TRANSFORMATION AND TRACING LANGUAGE

Relationship between model elements and a pattern instance is not established within the UML. Although there is the *Package* element defined within the UML, its purpose is not the same as "the pattern instance". Also, transformation application and imposing constraints on involved pattern instances must leave some trail. Creation of a Transformation and Tracing Model (TTM), either

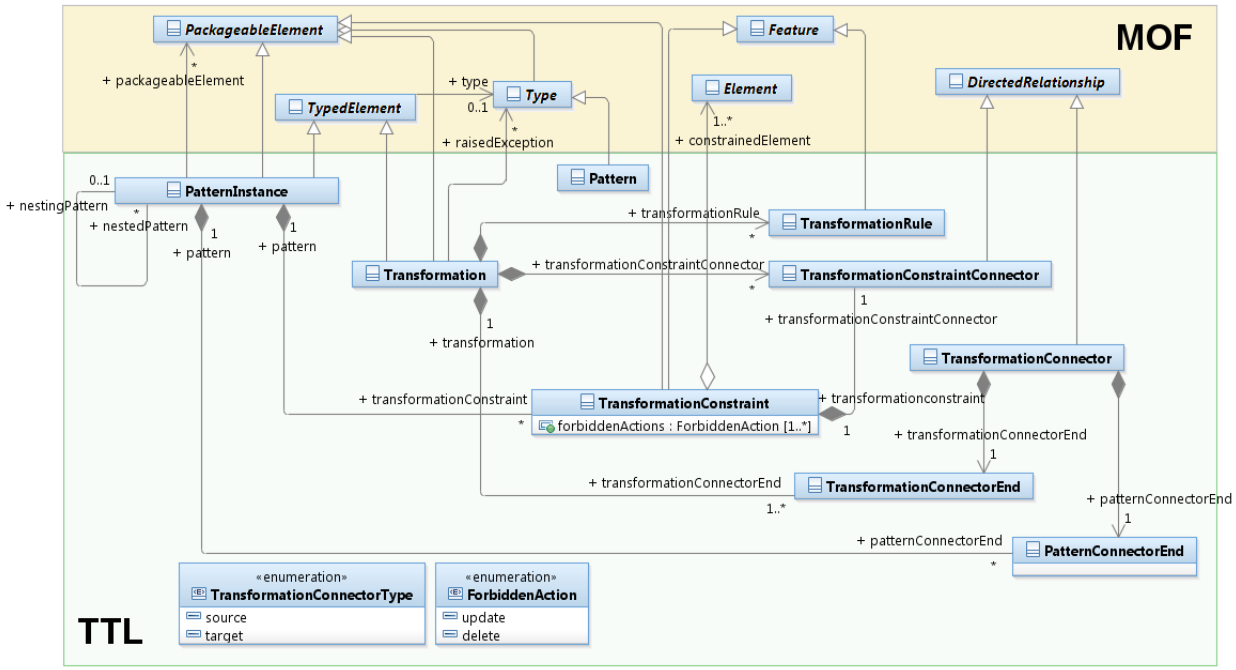


Figure 4. The Transformation and Tracing Language

automatically or manually, can help to resolve before mentioned issues. Every time a new pattern instance is created, a new model element is added into TTM representing this pattern instance. All model elements belonging to this pattern instance are automatically bound to it. It can be the result of the transformation, or it can be done manually. In both cases, the modeling tool must have capabilities for it. Also, each time when a transformation is used, this transformation is added to TTM including all relationships between pattern instances and used transformation. Each time a transformation is used, and this transformation is imposing constraints on involved pattern instances, these constraints are added to pattern instances in TTM and related to the transformation that created them. In order to do this kind of model, a Transformation and Tracing Language (TTL) must be defined. The UML and the TTL must be compatible, meaning that they must have a common M0 ancestor [28]. Therefore, the TTL must be a MOF metamodel. An overview of the TTL is presented in Figure 4.

The TTL is having the following elements:

1. *Pattern* - A pattern type. Allows classification of pattern instances.
2. *PatternInstance* - An element similar to the UML *Package* element. Represents a container for model elements. This element is defined by its name and type. Pattern type (or class) can be very helpful when constructing transformation rules, and it can impact the transformation applicability since transformations can be applied to the pattern instances of specific types.
3. *Transformation* - An element defined by its name and type, representing applied transformation, defined in (6) and (7). It contains transformation rules used in the transformation, here represented by the element *TransformationRule*. The transformation must be connected to a set of source and target pattern instances, being connected to at least one target pattern instance. Connector direction is determined by the *TransformationConnectorType* enumeration.
4. *TransformationConnector*, *TransformationConnectorEnd*, *PatternConnectorEnd* - A connector is a directed relationship between a pattern instance and a transformation. Connector direction must have a visual notation. If the connector is directed from the pattern instance to the transformation, it represents the source pattern instance in the context of the transformation. If the connector is directed from the transformation to the pattern instance, it represents the target pattern instance in the context of the transformation. Connector end elements represent the point of touch between the connector and the pattern instance, or the connector and the transformation.
5. *TransformationConstraint* - An element defined by its name, representing a constraint on members of a pattern instance imposed by used transformation. At least one model element in the pattern instance needs to be constrained. Also, a transformation must contain a set of forbidden actions, i.e., actions on constrained model elements that must be prevented by a tool. This element is contained by the pattern

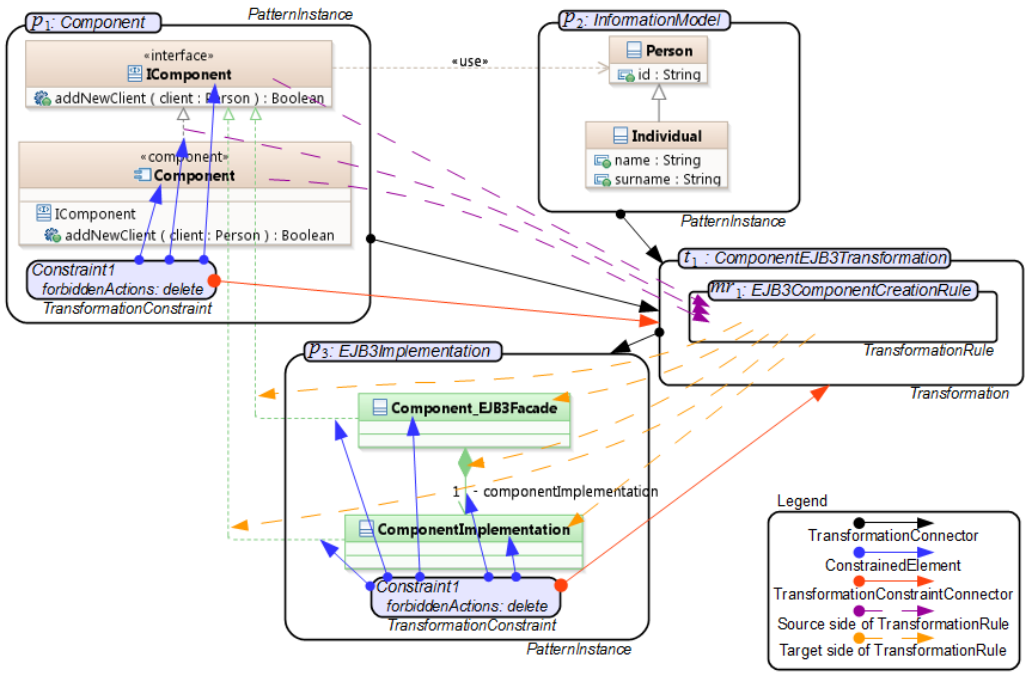


Figure 5. Example of a Transformation and Tracing Model

instance, and connected to the transformation responsible for the creation of the constraint. This element is the result of the transformation, and can be used to validate the pattern instance correctness and completeness.

6. *TransformationConstraintConnector* - A relationship between resulting constraint and the transformation that created it, directed from the transformation to the constraint. Each constraint can be imposed by only one transformation, but one transformation can impose multiple constraints within multiple pattern instances.

In the TTM example in Figure 5, model elements in pattern instances p_1 and p_2 were created before t_1 was applied. We can say that pattern instances p_1 and p_2 were designed manually. Model elements in the pattern instance p_3 are produced by the transformation t_1. Actions taken during an information system design are automatically stored to a TTM for multiple purposes: preserving correctness and completeness of the modeling space, reconstruction of activities in the design process, and analysis of the resulting design work.

VI. Design Practice

The definition of the term "design practice" is given in [2]. A common situation is having to explain to designers what is the preferred design practice, and how an information system design should look like? The answer to this question is also the answer to the design approach classification framework given in [2].

Many companies have well established design practices, from the methodology, project activities, and modeling point of view. The selection of architectures, technology, and practical experience gives a company starting point in the information system design. The idea is to take this experience, put it into the modeling library in the form of

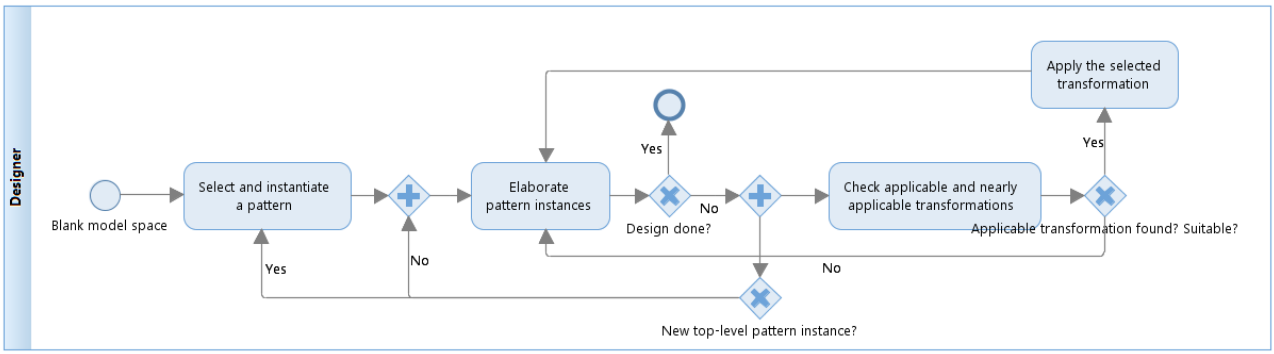


Figure 6. The proposed method

patterns and transformations, i.e., create a design knowledge system.

The way of applying this knowledge is very important as well. The design practice method proposed in this article, and outlined in Figure 6, consists of the following elements:

- Task 1: Selection and instantiation of appropriate pattern from the modeling library.
- Task 2: Elaboration of pattern instances.
- Task 3: Checking transformations from the modeling library that are applicable or nearly applicable onto elaborated pattern instances. If such transformation cannot be found, return to elaboration in task 2.
- Task 4: Transformation of the pattern instance by selecting applicable transformation from the modeling library. Continue on task 2 with newly created pattern instances.

Using the process in Figure 6 will create a pattern sequence. Depending on patterns and transformations in the modeling library, a big set of potential pattern sequences can be generated. Giving guidance to designers on a project means selecting appropriate pattern sequences from the set of all potential pattern sequences. By selecting appropriate pattern sequences, the design process is directed into the desired direction and outcome.

A. *Guidance given through the modeling library*

The modeling library is comprised of patterns and transformations. Since a transformation binds two pattern instances together (as described in Section III), selection of a transformation imposes a selection of involved patterns. Similarly, a selection of patterns imposes a selection of potentially applicable transformations.

Applicability and the measure of transformation applicability are important transformation features that can be used to form a pattern sequence. A designer can elaborate a model or a pattern instance, and occasionally check for transformations that are applicable to the model or pattern instance he is working on. If there is no transformation currently applicable, the designer can check transformations that are nearly applicable, and the gap that needs to be closed in the model or the pattern instance in order for this nearly applicable transformation to become applicable. Of course, many designers have enough experience to know which transformation would need to be used next, even before modeling of the pattern instance is finished. If there is a problem with selected transformation, and transformation rules in the transformation are not correct, meaning that the transformation will never become applicable, this particular transformation can be changed as part of the design practice evolution.

Giving guidance means selecting transformations from the modeling library that will be used in the project. A design lead can manage the set of allowed transformations for the project, limiting designer's choice of applicable or nearly applicable transformations. For example, the architectural decision will influence the choice of transformations for the project. Similarly, the design lead can manage a set of allowed patterns that are going to form the pattern sequences in the project, and by doing that implicitly to select a set of allowed transformations.

B. *Guidance given through a model*

More specific guidance can be given through a specific model that predetermines patterns and transformations used in the information system design process. Such model is created *a priori*, before the start of the design activities. Creation of the guidance model is an ongoing activity through the whole project. The TTL can be used for this purpose. This model must represent a selection of allowed pattern types and related transformations. Such model can be used by a designer to check guidance, or directly by a modeling tool for selection of allowed transformation list for particular pattern type. It is the same approach as in the previous section, with additional visualization of selected design practice for the project.

VII. EXAMPLE: BUSINESS PROCESS ORIENTED SYSTEM

The example in Figure 3 is business process oriented. The common name for this kind of system is Business Process Management (BPM) System. In this section, a detailed walk through for the example in Figure 3 is given. The first step is to create a business process model. Such business process model can be done using BPMN [24].

In this case, a very simple business process is modeled from scratch. A new pattern instance p_1 is created and placed in the TTM. When modeling, the model elements of the business process are associated with the pattern instance p_1. The business process contains one human task having the

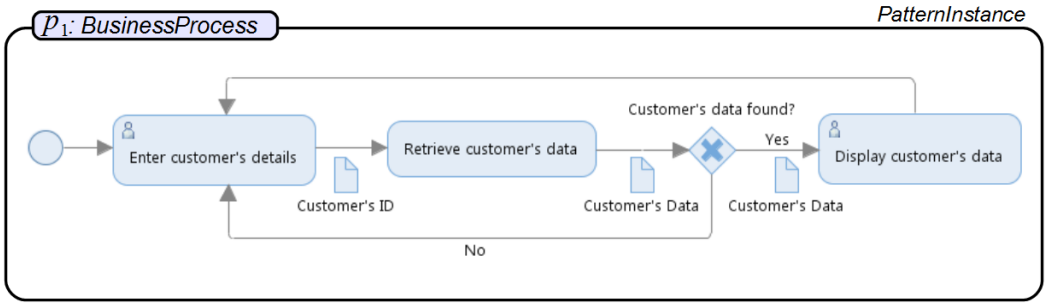


Figure 7. The business process

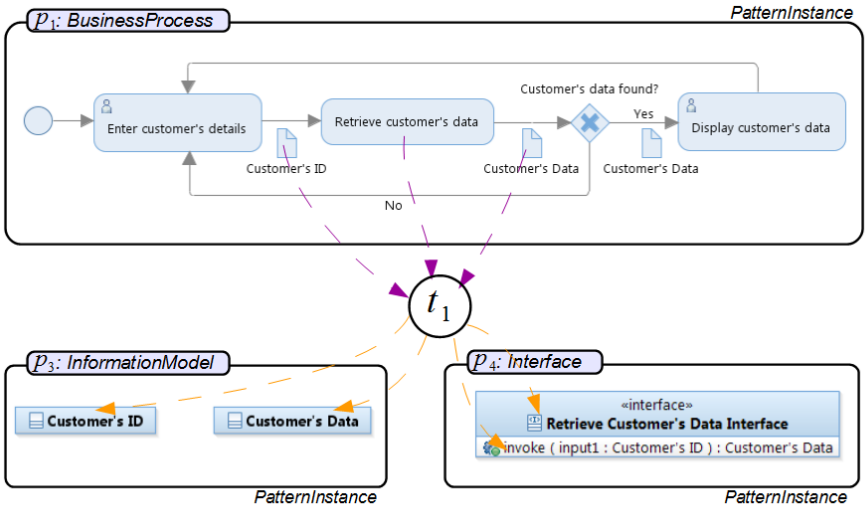


Figure 8. Transformation of the task to entities and the interface

user interface, where a user must enter customer's identification data, such as "social id", "account number", or something else. The customer's identification is then sent to the automatic task named "Retrieve customer's data", which invokes an information service that finds and read the data. If the identified customer cannot be found in the system, a null is returned back from the invoked service. Returned data are then tested, and if customer's data exists, it is displayed in the human task "Display customer's data".

Each automatic task contains a signature that involves input parameters and output results. These details must be observed on a correctly modeled business process, as in Figure 7. This signature is used for transforming this automatic task into a number of pattern instances, containing entities and interfaces needed for building the service that will be invoked by this task. Each business process can have more than one automatic task. This means that one transformation from the modeling library can be applied more than once per one pattern instance. A designer must be presented with the list of applicable transformations along with all details, including model elements in source pattern instances that can be used in the transformation. In case of a business process that contains more than one automatic task, an applicable transformation can be applied to each automatic task.

Figure 8 presents a transformation from the automatic task in the business process into a set of entities and an interface. After applying the transformation t_1, the pattern instance p_3 is created. Along with the pattern instance, model elements representing entities of two business objects constituting information flow of the transformed task are created. The transformation t_1 must create constraints that will prevent deleting business objects, the task in the business process, and both of the entities in the pattern instance p_3.

The transformation t_1 is also responsible for creation of the pattern instance p_4, which consists of model elements that represents the interface of the task in the business process: one operation receiving the input business object as the input parameter, and returning the output business object as the result. The transformation t_1 must create constraints that will prevent deleting the involved model elements in pattern instances p_3 and p_4. Another constraint that will prevent direct updating the operation on the interface must be created by the transformation t_1 as well, because updates on transformation's target pattern instances must be result of the elaboration of the source pattern instances. It is worth noticing that an operation and comprising parameters are, according to the MOF, not the same model elements. While an operation can be constrained, comprised parameters can be updated by the transformation if there are changes on business objects in the business process.

The next step is to elaborate the pattern instance p_3. As presented in Figure 9, a designer is filling additional details for entities in the pattern instance p_3. These details include attributes and their types for each entity.

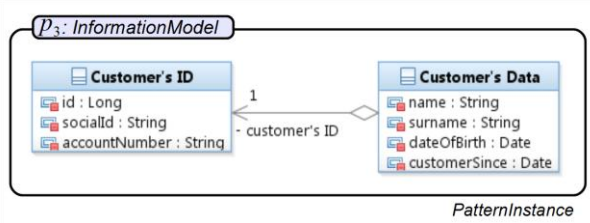


Figure 9. Elaborated entities

At the same time, a lead IT architect is defining the architecture of the information system. It is very important that the architecture is a part of the modeling space, so that it can be used by the proposed method for selection of allowed transformations. This way, the architecture guides the design process as well.

The architecture presented in Figure 10 defines a couple of very important elements for the selection of applicable

transformations. First, JAX-WS2 will be used for invoking the service from the automated task in the business process. Second, the service will be deployed on Java Enterprise Edition application server. Third, entities will become tables in the DB2 database.

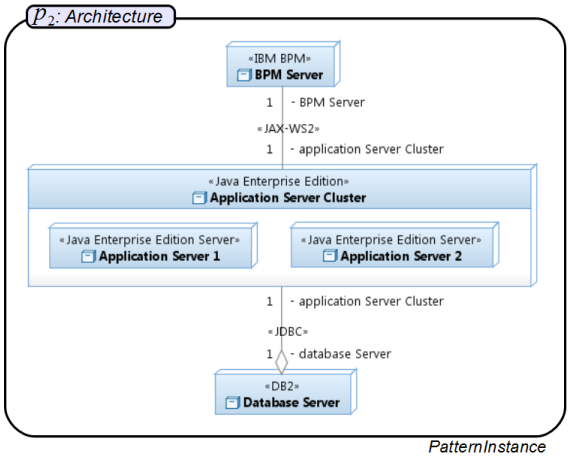


Figure 10. The architecture

Figure 10 is one pattern instance. This would suggest that the whole architecture is placed in the single pattern instance container. In fact, Figure 10 represents a typical business process oriented architectural pattern specified for the concrete environment and products.

So far, the design is still not platform specific. Pattern instances p_3 and p_4 can be transformed into a component.

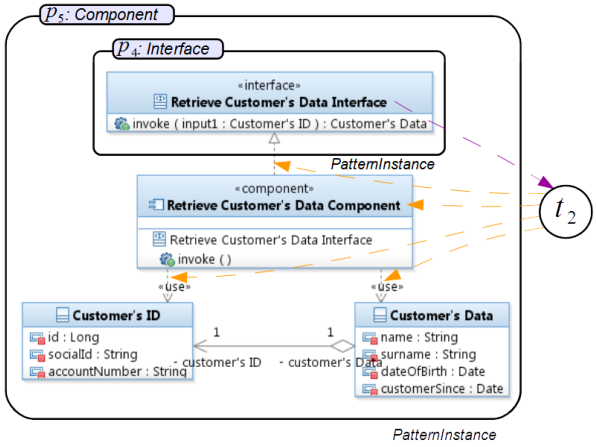


Figure 11. The component

Such transformation must create a component, and define that it is realized using already created interface, as in Figure 11. For reference purpose, transformation t_2 adds dependencies between the created component and entities contained in the realized interface.

After this, work on the platform specific design can begin. The next step is to elaborate additional details in the pattern instance p_3 in order to make transformation t_3 applicable.

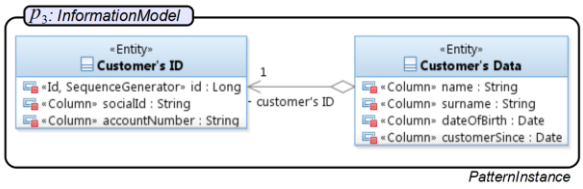


Figure 12. Entities with JPA stereotypes

During elaboration, entities in the pattern instance p_3 are enriched with JPA related stereotypes, as in Figure 12. Since transformation t_3 takes in consideration model elements marked with *Entity* stereotype, this elaboration is needed in order to make the transformation applicable. During the elaboration of entities, a designer must also define additional properties for applied stereotypes. For example, a *Column* stereotype has a set of very important properties that need to be defined, and can be used by transformations that will be applied next. Figure 13 presents a set of properties for the *Entity* stereotype.

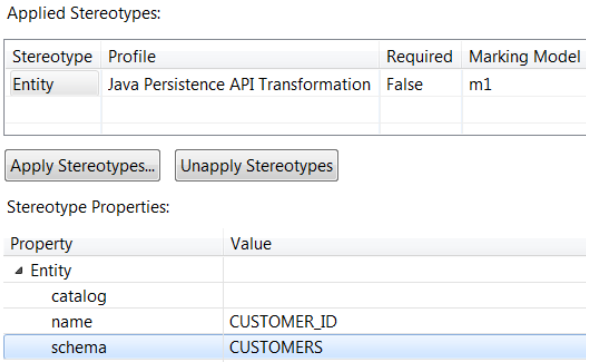


Figure 13. Entity stereotype properties

However, this elaboration makes the pattern instance p_3 more platform specific than platform independent. Although there is still no precise definition about the concrete database that will be used, this information can be found in the architecture contained in the pattern instance p_2.

Figure 14 is the result of applying transformation t_3 to the pattern instances p_2 and p_3. The transformation takes only entities marked with *Entity* stereotypes. All properties of applied stereotypes are used in the transformation. Also, the transformation matches the database node in the architecture pattern instance p_2. The result of the transformation is the pattern instance p_6 that comprises a number of DB2 specific model elements, representing database tables of entities.

In case of having an Oracle database node in the architecture, another transformation would become applicable, which would create Oracle specific model elements in the pattern instance p_6.

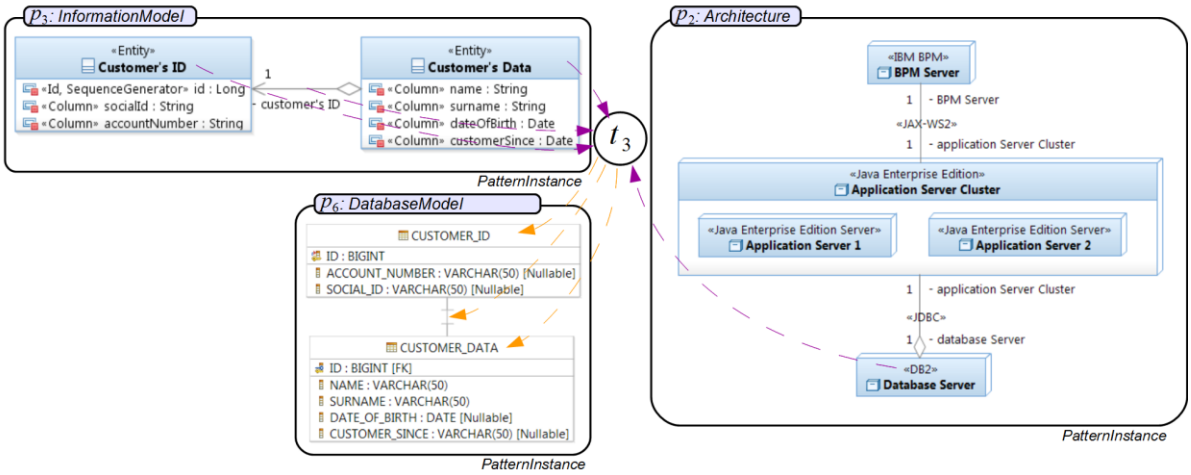


Figure 14. Entities with JPA stereotypes

The pattern instance p_6 is a true example of a platform specific design. Now that the information part of the design is done, designing component details is the next step.

The component implementation presented in Figure 15 is a result of two steps. The first step is applying transformation t_4 for creation of a component implementation. This transformation is applicable only when *JAX-WS2* and *Java Enterprise Edition Server* stereotypes are found in the architecture pattern instance. It creates a class with appropriate stereotypes for further transformations, an interface realization between the newly created class and the component's interface, and a relationship that marks the newly created class as an instance of the component. However, transformation t_4 did not create any specific implementation details. Everything that was created is the class will be eventually transformed into a JAX-WS2 web service provider.

The second step is applying transformation t_5 that adds implementation details to the service provider created in the previous step. Again, a designer can have a number of transformations at disposal that can look for certain model elements in the existing pattern instances. In this case, transformation t_5 takes the signature of the *invoke* method, parameter types and creates a method in a new JPA reading

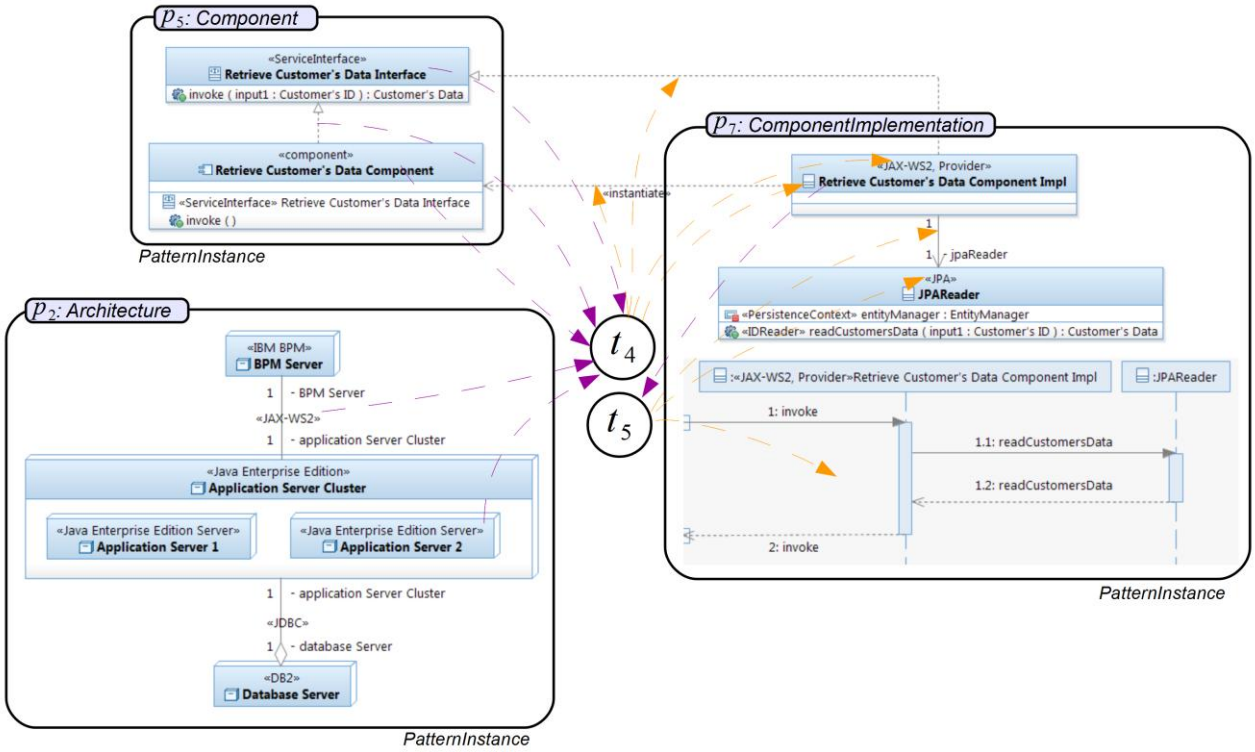


Figure 15. The component implementation

helper class. This method is marked with the *IDReader* stereotype, which can be used later to transform this method into a code snippet. Transformation t_5 also creates an association between the component's implementation and the newly created JPA reading helper class. A collaboration sequence between the component's implementation class and the JPA reading helper class is created as well. This collaboration sequence can be used in creating the code later on.

As mentioned in [26], transformations from model to code need to be treated slightly differently. Template based approach is suitable for this example.

Source Code 1: Transformed JPA entity

```
@Table(name = "CUSTOMER_ID", schema = "CUSTOMERS")
@Entity
public class CustomersID implements Serializable {

   private static final long serialVersionUID = 0;

   public CustomersID() {}

   @Id
   private Long id;

   @Column(nullable = false, columnDefinition =
        "SOCIAL_ID", length = 50)
   private String socialId;

   @Column(nullable = false, columnDefinition =
        "ACCOUNT_NUMBER", length = 50)
   private String accountNumber;

   public Long getId() {
      return id;
   }

   public void setId(Long id) {
      this.id = id;
   }

   public String getSocialId() {
      return socialId;
   }

   public void setSocialId(String socialId) {
      this.socialId = socialId;
   }

   public String getAccountNumber() {
      return accountNumber;
   }

   public void setAccountNumber(String accountNumber) {
      this.accountNumber = accountNumber;
   }
}
```

The listing in Source Code 1 is the final result of transforming from entities in the pattern instance p_3. Such transformations can be applicable only on *InformationModel* pattern instances, i.e., including only model elements that are contained in specific pattern instances.

Using pattern instance p_6 and the applicable transformation, the following SQL script is generated.

SQL Script 1: Database DDL script

```
CREATE SCHEMA "CUSTOMER";

CREATE TABLE "CUSTOMER"."CUSTOMER_DATA" (
                "NAME" VARCHAR(50) NOT NULL,
                "SURNAME" VARCHAR(50) NOT NULL,
                "DATE_OF_BIRTH" DATE,
                "CUSTOMER_SINCE" VARCHAR(50),
                "ID" BIGINT NOT NULL
        )
        DATA CAPTURE NONE;

CREATE TABLE "CUSTOMER"."CUSTOMER_ID" (
                "ID" BIGINT NOT NULL GENERATED BY
DEFAULT AS IDENTITY ( START WITH 1 INCREMENT BY 1
MINVALUE 1 MAXVALUE 9223372036854775807 NO CYCLE CACHE
20),
                "ACCOUNT_NUMBER" VARCHAR(50),
                "SOCIAL_ID" VARCHAR(50)
        )
        DATA CAPTURE NONE;

ALTER TABLE "CUSTOMER"."CUSTOMER_DATA" ADD CONSTRAINT
"CUSTOMER_DATA_PK" PRIMARY KEY
        ("ID");

ALTER TABLE "CUSTOMER"."CUSTOMER_ID" ADD CONSTRAINT
"CUSTOMERS_ID_PK" PRIMARY KEY
        ("ID");

ALTER TABLE "CUSTOMER"."CUSTOMER_DATA" ADD CONSTRAINT
"CUSTOMER_DATA_CUSTOMER_ID_FK" FOREIGN KEY
        ("ID")
        REFERENCES "CUSTOMER"."CUSTOMER_ID"
        ("ID")
        ON DELETE CASCADE;
```

Finally, pattern instances p_5 and p_7 can be transformed into the web service that can be called from the task in the business process.

Source Code 2: The web service interface

```
@WebService(targetNamespace="customer")
public interface RetrieveCustomersDataInterface {
   @WebMethod
   public CustomersData invoke(CustomersID input1);
}
```

Source Code 3: The web service

```
@WebService(targetNamespace="customer")
public class RetrieveCustomersDataComponentImpl
implements RetrieveCustomersDataInterface {
   private JPAReader jpaReader;

   public CustomersData invoke(CustomersID input1) {
      // begin-user-code
      return jpaReader.readCustomersData(input1);
      // end-user-code
   }
}
```

Source Code 4: The JPA reader

```
public class JPAReader {
   @PersistenceContext
   private EntityManager entityManager;
   // TODO Finish instancing entity manager

   public CustomersData readCustomersData(CustomersID
        input1) {
      // begin-user-code
      Query q=null;
      if(input1!=null && input1.getId()!=null) {
         q=entityManager.createQuery("select obj from
                CustomersData obj where
                obj.customersID.id = :id");
         q.setParameter("id", input1.getId());
      } else if(input1!=null &&
         input1.getSocialId()!=null) {
         q=entityManager.createQuery("select obj from
                CustomersData obj where
                obj.customersID.socialId = :socialId");
         q.setParameter("socialId",
                input1.getSocialId());
```

```
        } else if(input1!=null &&
           input1.getAccountNumber()!=null) {
           q=entityManager.createQuery("select obj from
                    CustomersData obj where
                    obj.customersID.accountNumber =
                    :accountNumber");
           q.setParameter("accountNumber",
                    input1.getAccountNumber());
        }
        if(q!=null) {
           return (CustomersData)q.getSingleResult();
        }
        return null;
        // end-user-code
    }
}
```

The call between the component's implementation and the JPA helper class is transformed from the collaboration sequence. All additional details in the code, such as annotations, are added from stereotype information. Stereotypes on the JPA helper class help to select an appropriate template for the code.

VIII. Conclusion and Future Work

We have demonstrated that even such small example can be full of details and rules, enforcing us to use specific model elements, stereotypes, and patterns. It is obvious that patterns are not just a couple of documented ways of solving problems, which can be found in the books. It is everything that we want to use to repeat our solutions. Patterns are a good start for defining our designing practice.

MDA has two major practical problems: designers have too much freedom while creating the information system design so that the transformation scope can become very ambiguous. Usage of a pattern as the main building element for the information system design is a well known approach. In the context of this article, design of an information system is done block by block by reusing patterns, allowing a design lead to choose blocks to be used. Such approach allows a design team to use past positive experience to select or define best patterns for the information system they are designing. This approach also helps to build pattern sequences that can fit into a design and development methodology used for the project.

The novelty introduced in this article is the way of building pattern sequences through use of transformation, an approach typically used in the MDA. Applicability and the measure of applicability are very important features of the transformation definition, given in this article. They enable controlled application of transformations, which represents guidance for the design team. They also represent a way how new designers can learn the established design practice.

Of course, designers are still free to model according to their preferences, as long as they are within boundaries imposed by the proposed method, which is assured by an optional part of each transformation helping team to keep model elements of bound pattern instances synchronized. The bidirectionality feature of the transformation helps to reflect changes in both directions. Chains of pattern instances can be easily updated through transformations used to form a chain. Since a pattern instance is supposed to have smaller scope than a model, keeping several pattern instances synchronized during elaboration should be much easier than with big models.

The proposed method successfully answers challenges introduced in Section I. The modeling library contains design knowledge, and offers a selection of the design approach based on the current context in the modeling space. The article also successfully answers question of transformation reusability. The result of all these improvements is a higher quality of models comprising the design of an information system.

Current modeling tools are introducing a high level of automation. This automation is mostly related to elements of the modeling languages supported by a modeling tool. Changing the modeling tool behavior to follow the model in a modeling space is needed feature.

The TTL defined in this article can be extended with elements for interaction with modeling tool, model analysis capabilities, and model quality assessment. Interaction between a TTM and a modeling tool can be extended with modeling events, allowing a design lead to define modeling tool actions associated with patterns and transformations. For example, a TTM can include an event handler on a pattern that can be triggered by the modeling tool when a new subcomponent is added into a pattern instance. The event handler initiates execution of a specific transformation that automatically adds interface and interface realization relationship for this newly added subcomponent.

References

[1] D. Krleža and K. Fertalj, "A method for situational and guided information system design," Proceedings of the 6th International Conference on Pervasive Patterns and Applications, IARIA, May 2014, pp. 70-78.

[2] P. Ralph and Y. Wand, "A proposal for a formal definition of the design concept," in Design requirements engineering: A ten-year perspective, Springer Berlin Heidelberg, vol. 14, pp. 103-136, 2009, doi: 10.1007/978-3-540-92966-6_6.

[3] Object Management Group, "MDA guide, version 1.0.1, 2003". Available from http://www.omg.org/cgi-bin/doc?omg/03-06-01.pdf 2014.11.15

[4] M.F. Gholami and R. Ramsin, "Strategies for improving MDA-based development processes," Proceedings of the 2010 International Conference on Intelligent Systems, Modelling and Simulation (ISMS), IEEE, Jan. 2010, pp. 152-157, doi: 10.1109/ISMS.2010.38.

[5] L. Osterweil, "Software processes are software too, revisited: an invited talk on the most influential paper of ICSE 9," Proceedings of the 19th international conference on Software engineering ICSE '97, ACM, May 1997, pp. 540-548, doi: 10.1145/253228.253440.

[6] C. F. J. Lange and M. R. V. Chaudron, "Managing model quality in UML-based software development," 13th IEEE International Workshop on Software Technology and Engineering Practice, IEEE, Sep. 2005, pp. 7-16, doi: 10.1109/STEP.2005.16.

[7] F. Chitforoush, M. Yazdandoost, and R. Ramsin, "Methodology support for the model driven architecture," Proceedings of the 14th Asia-Pacific Software Engineering Conference, IEEE, Dec. 2007, pp. 454-461, doi: 10.1109/ASPEC.2007.58.

[8] I. Jacobson, G. Booch, and J. E. Rumbaugh, The unified software development process - the complete guide to the unified process from the original designers, Addison-Wesley, 1999.

[9] P. Kroll and P. Kruchten, The rational unified process made easy: a practitioner's guide to the RUP, Addison-Wesley, 2003.

[10] J. Rumbaugh, I. Jacobson, and G. Booch, The Unified Modeling Language Reference Manual, Addison-Wesley, 1999.

[11] C. Alexander, S. Ishikawa, M. Silverstein, M. Jacobson, I.Fiksdahl-King, and S. Angel, A pattern language: towns, buildings, constructions, Oxford University Press, 1977.

[12] C. Fehling, F. Leymann, R. Mietzner, and W. Schupeck, "A collection of patterns for cloud types, cloud service models, and cloud-based application architectures". Available from http://www.cloudcomputingpatterns.org 2014.11.15

[13] C. Fehling, F. Leymann, R. Retter, D. Schumm, and W. Schupeck, "An architectural pattern language of cloud-based applications," Proceedings of the 18th Conference on Pattern Languages of Programs, ACM, Oct. 2011, pp. 2, doi: 10.1145/2578903.2579140

[14] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, Design patterns: Elements of reusable object-oriented software, 28th ed., Addison-Wesley, 2004.

[15] G. Hohpe and B. Woolf, Enterprise Integration Patterns, Addison Wesley, 2004.

[16] Object Management Group, "Meta object facility (MOF) core specification, version 2.4.2, 2014". Available from http://www.omg.org/spec/MOF/2.4.2/PDF/ 2014.11.15

[17] S.D. Frankel, Model driven architecture: applying MDA to enterprise computing, Wiley Publishing, 2003.

[18] M. Gupta, R. Singh Rao, and A. Kumar Tripathi, "Design pattern detection using inexact graph matching," 2010 International Conference on Communication and Computational Intelligence, IEEE, Dec. 2010, pp. 211-217.

[19] N. H. Pham, H. A. Nguyen, T. T. Nguyen, J. M. Al-Kofahi, and T. N. Nguyen, "Complete and accurate clone detection in graph-based models," Proceedings of the 31st International Conference on Software Engineering, IEEE, May 2009, pp. 276-286, doi: 10.1109/ICSE.2009.5070528.

[20] M. Falkenthal, J. Barzen, U. Breitenbücher, C. Fehling, and F. Leymann, "From pattern languages to solution implementations," Proceedings of the 6th International Conference on Pervasive Patterns and Applications, IARIA, May 2014, pp. 12-21.

[21] D.R. Stevenson, J.R. Abbott, J.M. Fischer, S.E. Schneider, B.K. Roberts, M.C. Andrews, D.J. Ruest, S.K. Gardner, and C.D. Maguire, "Pattern implementation technique," U.S. Patent 8 661 405, Feb. 25, 2014.

[22] R. Porter, J. O. Coplien, and T. Winn, "Sequences as a basis for pattern language composition," Science of Computer Programming, Elsevier, vol. 56, pp. 231–249, Apr. 2005.

[23] W. Hasselbring, "Component-based software engineering," Handbook of Software Engineering and Knowledge Engineering, World Scientific Publishing, vol. 2, pp. 289-305, 2002, doi: 10.1142/9789812389701_0013.

[24] Object Management Group, "Business process model notation, version 2.0.2, 2013". Available from http://www.omg.org/spec/BPMN/2.0.2/PDF/ 2014.11.15

[25] L. Grunske, L. Geiger, A. Zündorf, N. Van Eetvelde, P. Van Gorp, and D. Varro, "Using graph transformation for practical model-driven software engineering," in Model-driven Software Development, Springer Berlin Heidelberg, pp. 91-117, 2005, doi: 10.1007/3-540-28554-7_5.

[26] K. Czarnecki and S. Helsen, "Classification of model transformation approaches," Proceedings of the 2nd OOPSLA Workshop on Generative Techniques in the Context of the Model Driven Architecture, vol. 45, no. 3, Oct. 2003, pp. 1-17.

[27] Object Management Group, "Meta object facility (MOF) 2.0 query/view/transformation (QVT), version 1.1, 2011". Available from http://www.omg.org/spec/QVT/1.1/PDF/ 2014.11.15

[28] A. G. Kleppe, J. B. Warmer, and W. Bast, MDA explained, the model driven architecture: Practice and promise, Addison-Wesley, 2003.

[29] F. Jouault and I. Kurtev, "Transforming models with ATL," Proceedings of the MoDELS 2005 Conference, Springer Berlin Heidelberg, Oct. 2005, pp. 128-138, doi: 10.1007/11663430_14.

[30] A. Van Gelder, K. A. Ross, and J. S. Schlipf, "The well-founded semantics for general logic programs," Journal of the ACM (JACM), ACM, vol. 38, no. 3, pp. 619-649, Jul. 1991, doi: 10.1145/116825.116838.

Architectural Design Considerations for Context-Aware Support in RECON Intelligence Analysis

Alexis Morris*, William Ross*, Mihaela Ulieru†, Daniel Lafond‡, René Proulx‡, and Alexandre Bergeron-Guyard§

*Faculty of Computer Science, University of New Brunswick, Fredericton, Canada
{alexis.morris, william.ross}@unb.ca
†School of Information Technology, Carleton University, Ottawa, Canada
{mihaela}@theimpactinstitute.org
‡Thales Research and Technology Canada, Quebec City, Canada
{daniel.lafond, rene.proulx}@ca.thalesgroup.com
§Defence Research and Development Canada (Valcartier), Quebec City, Canada
{alexandre.bergeron-guyard}@drdc-rddc.gc.ca

***Abstract*—The REcommending Cases based on cONtext (RECON) system is a prototype adaptive technology designed to support intelligence analysts in overcoming the problem of cognitive overload. Its central objective is to assist these analysts during the collection, processing, and analysis phases of the intelligence cycle through sense-making of both explicit and implicit contextual information. RECON combines machine learning, text-analysis, brain-computer interfaces, and simulation to create an innovative case-based recommendation capability. In developing RECON, multiple considerations have been explored based on key human-computer interaction dilemmas that emerge when designing joint-cognitive systems endowed with an adaptive capacity. Herein, eight architectural design considerations are discussed, related to human-modelling, human-machine interaction, and human-machine synergy, which have impacted the system development. The central RECON architecture and its components are also presented, including a context-sensitive cognitive model based on COCOM. This work aims to provide these core architectural components and their design considerations as a contribution toward aiding developers in designing, customizing, and improving future adaptive context-management systems.**

Keywords–adaptive systems; context-awareness; human factors; human-computer interaction; brain-computer interfaces.

I. INTRODUCTION

The development of adaptive software systems and infrastructures that are situationally responsive and human-centred remains an attractive area of research, as technologies progress and people continue to work with more data as part of their routine tasks. The improvement of human-centred technologies involves a synergy of both human and machine in order to address the dynamics of unfolding situations; hence, systems that are both dynamic and responsive are required. These dynamic systems are inherently open, interacting with the environment of the organization to carry out its goals, which are often time-sensitive, as in the case of real-time information systems, or even critical, as in the case of emergency response. Moreover, the problem of information overload is becoming increasingly evident in the world, as people become more connected with technology, and this trend is only expected to continue with the proliferation of "big data."

Having technology that can adapt to the varied needs of the user—be they task-related or cognitive-related needs—and having systems that can incorporate humans-in-the-loop to sift through large volumes of data in order to effectively gather and assess information offer direction toward a possible solution to the problem. These point directly to a context-sensitive approach for achieving human-machine synergy, where the combined results of the human and software system working together is greater than the result of any one component working in isolation. The development of such systems requires design considerations that are both human-computer-interface (HCI)-focused and context-based, as discussed in the authors' previous work from ADAPTIVE 2014 [1], which highlights core HCI dilemmas for context-aware support in the intelligence analysis domain.

Adaptation in human-machine systems is challenging, as it requires significant information monitoring. The human must monitor incoming information in order to determine appropriate decisions and response actions, and the technological system must monitor user-context information in order to adapt to the user and perform functions in a dynamic environment. Also, human-machine systems involve the often-complex interplay of human and technological components as interconnected actors sharing a common goal. To be agile, both the human-in-the-loop and the technological system must be in sync with the speed, scope, and context of real-world dynamics, as interactions in a complex real-world situation require corresponding complexity in adaptive systems [2]. However, it is known that these systems, which combine the human-social and technological dimensions, can often become out-of-sync in fast-paced situations where human decision-makers routinely require actions that are outside of the design scope of the technological systems on that they depend [3]. There is a need, therefore, for technology to support users in varied situations that are inherently human-centred, and such a practical adaptive system should enable the user to have balanced access to the most relevant information available (especially in information-centered domains), while also providing this information in a timely manner, in sync with the user's total context.

The current paper extends the authors' previous effort in [1] with a more comprehensive presentation of the architectural design considerations for the developed RECON (REcommending Cases based on cONtext) architecture, which will be presented in detail. Moreover, as these strongly influence the resulting functionality of the developed software system,

architectural design considerations should be carefully and explicitly examined and conscientiously applied. Herein, eight key considerations encountered during the course of design and early implementation will be presented and critically discussed. This work promotes these design considerations, which have been accounted for in the development of RECON, as an important step towards the development of future architectures for adaptive, context-aware management, and the expected audience for such design considerations are those involved in the information analysis domain, where cognitive overload is prevalent. The developed RECON system will be used as a case study in how to apply these considerations.

The remainder of this paper is organized as follows. Section II highlights the intelligence-analysis domain, including the problem of cognitive overload and context-awareness. Section III outlines the use case and detailed architecture for RECON. Section IV introduces relevant design considerations for context-aware systems development, along with a taxonomy of such considerations, while Section V presents how the introduced design considerations are applied to the RECON system implementation. Next, Section VI highlights related work on architectural design considerations and compares these with those applied in the RECON case. Lastly, Section VII concludes the paper and offers potential avenues for future work.

II. Problem Domain

In this section, the problem domain is described in more detail, along with the challenge of cognitive overload and a promising path toward a potential solution involving context awareness.

A. Intelligence Analysis Domain

To motivate the need for context-aware architectures in the intelligence domain, it is important to highlight the typical information cycle and its effect on the role of the intelligence analyst, as a key player, interested in sense-making and accurate projections for multiple situations that are often time-sensitive and multi-faceted. This intelligence cycle is defined as "the process of developing raw information into finished intelligence for policymakers to use in decision-making and action" [4]. The intelligence cycle encompasses many sense-making tasks that the intelligence analyst must accomplish in an iterative fashion. Such tasks include: gathering relevant information, representing and organizing the information in a schematic way that will ease the analysis process, developing an understanding of the situation by subjecting the information to various hypotheses, and producing intelligence packages and recommendations for courses of action.

As described by Pirolli and Card [5], the overall process of sense-making is organized into two major loops of activities: (1) a *foraging loop* that involves processes aimed at seeking, searching, filtering, reading and extracting information, possibly into some schema [6]; and (2) a *sense-making loop* that involves iterative development of a mental model (a conceptualization) from the schema that best fits the evidence [7]. This process is illustrated in Figure 1.

The intelligence cycle, described in [5], is shown in Figure 2. This cycle includes activities involving planning and direction, collection of data, processing of data, analysis and production of resolutions and projections, and dissemination of

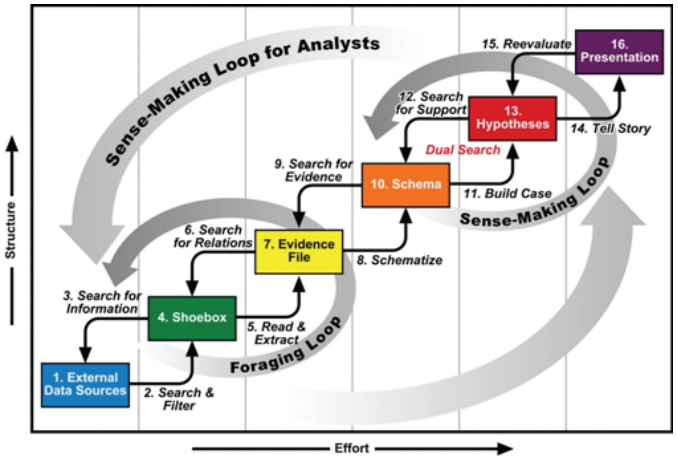


Figure 1. Notional model of sense-making (from [5]).

Figure 2. The intelligence cycle (adapted from [10]).

information to decision-makers. Here, the intelligence analyst performs the role of seeking out information for a set of unfolding situations, many of which may be dynamic and fast-changing, through collation of multiple documents.

While the day-to-day activities of the intelligence analyst are driven by this intelligence cycle, the analyst's activities are subjected to a number of contextual factors (e.g., psycho-physiological and environmental) that can severely impede intelligence analysis due to excessive workload, time pressure, and uncertainty [8]. However, the primary cause of concern is that of cognitive overload, which impacts the analyst's ability to effectively identify situationally-relevant information due to data overload (i.e., too much data to sift through) and/or cognitive limitations (i.e., too much complexity in the data for making immediate sense without assistive analytical tools) [9]. Together, these present a critical challenge to the development and success of advanced adaptive systems, where humans-in-the-loop must make sense of an ever-increasing inflow of data in order to perform their tasks.

B. Context Awareness

To manage the dynamics of real-world information monitoring and sense-making, there are many different contexts

that can be considered by an adaptive system. However, the challenge is in finding the "right" context so that the system, in turn, can act as an aid (rather than a hindrance) to the expert human user. As in [8], *context* is considered as anything that can be used to correctly identify the situation of a user. Context can be provided directly by the user or generated based on the user's actions, such as system tasks recently performed (based on system logs) and current location data (based on mobile global-positioning systems) [11]. Context can also refer to less concretized notions, such as describing users' psycho-physiological states, including their current cognitive mood and stress level. These can be obtained through active and passive sensing of users via bio-metric sensors, but can also be deduced from other sources such as camera monitoring of facial expressions [12].

The successful management of both kinds of context is important. Systems that are adaptive to the dynamics of a wide range of contexts can increasingly support properties favouring the "5 Rights" [13]—i.e., providing the right information to the right person in the right place, at the right time, and in the right way (e.g., based on the preferences of the user). Practical systems that are aware of users with this level of detail are rare, although context-awareness has been a research staple for the past decade [14]. However, such systems are becoming more tenable due to advances in technologies for unobtrusively monitoring users' psychological and physiological states, combined with the technological trends towards miniaturization and improved efficiencies in computational speed and memory costs. As a result, it is now possible to develop better adaptive human-machine systems, synergistically enhancing both human and machine intelligence.

This section outlined the domain of study, namely, intelligence analysis, and the problem under investigation, namely, reducing cognitive overload. It also highlighted the relevance of context-awareness as a promising solution area. These are examined further in the following section, which presents the five-layer, context-aware RECON architecture in detail.

III. RECON: Recommending Cases Based on Context

The RECON (REcommending Cases based on cONtext) system is a recent initiative aimed at providing a capability for intelligence analysts that takes into account their need for relevant information consumption in a time-sensitive environment. As part of Defence Research and Development Canada's iVAC (Intelligent Virtual Analyst Capability) project [15], RECON uses an adaptive-systems approach for information offloading and filtering to assist intelligence analysts. To this end, it focuses on the following three objectives, and in this section both the general use case and resulting architecture for RECON are described:

1) To support the analyst through appropriate visualization and selection of information based on user preferences and real-time brain-state information;
2) To enable the analyst to offload cognitive processing to the system for machine analysis and case-based recommendation; and
3) To alert the analyst through natural interfaces to relevant information based on context.

A. RECON Use Case

In conducting intelligence analysis, analysts must make sense of a variety of information sources (e.g., text documents, webpages, and supplementary GIS data sources). They assess this data according to specific goals and elements-of-interest as dictated by information stakeholders or other supervisors. They must also account for the severity of an unfolding situation, time projections (i.e., time to an event), and constraints in their deliberation strategies. As a result, analysts inevitably incur cognitive pressures, such as fatigue, attention loss, and stress, as they get closer to decision deadlines and perform longer work sessions. The RECON system aims to allow these analysts to offload aspects of this data collation and processing to its automated reasoner, by accepting analyst preferences and making appropriate document recommendations while adapting to real-time brain-monitoring data from the analyst. The general use case for RECON is outlined in Figure 3, along with a system overview diagram.

In this use case, an analyst (outfitted with a wearable and wireless *brain-computer interface (BCI) monitor*) performs the task of sense-making by reading through a number of text-based documents to identify patterns and trends that are applicable to the current situation-of-interest. The analyst logs into RECON using his or her profile information and defines the situation-of-interest in the form of *objectives*, namely a combination of entity and event relationships, through the use of the *HCI interface*. This interface further allows the analyst to input or update system settings and view documents, document recommendations, and scene notifications. A *scene* is an aspect of a situation that the analyst wishes to offload to the system; this can be defined using a combination of keywords, based on active objectives, parameter thresholds, and simulation configurations set by the analyst (e.g., the analyst may be interested in being notified when $> X$ documents are found containing a particular set of keywords). Simultaneously, the BCI headset performs monitoring and classification of the analyst's brain-waves to deduce states that indicate his or her psycho-physiological responses to the task at hand (e.g., whether the analyst is interested in the material within the document, or perhaps is experiencing the negative effects of cognitive overload).

During the course of the analyst's session with RECON, document recommendations are presented to the analyst as part of a document-notification interface, wherein an analyst may inspect each document. The documents listed in this interface are the result of a *recommendation algorithm*, which identifies each document's relevance to the analyst's objectives (defined using keywords and relationships). These documents are retrieved from available repositories, such as online news sites, social media, blogs, and document databases, and are gathered into RECON by a *data text analyzer* that uses existing text-analysis tools to add keyword metadata (used by the recommender) and perform sentiment analysis on the documents. The *context manager*, using its internal logic and the cognitive state of the analyst deduced by the BCI monitor, then determines when to notify the user about new documents and scene-threshold alerts. Lastly, a central *database* stores system-specific information, such as the objectives and preferences of analyst user accounts, enabling the system to continue working toward the defined goals even after the analyst has left for the day.

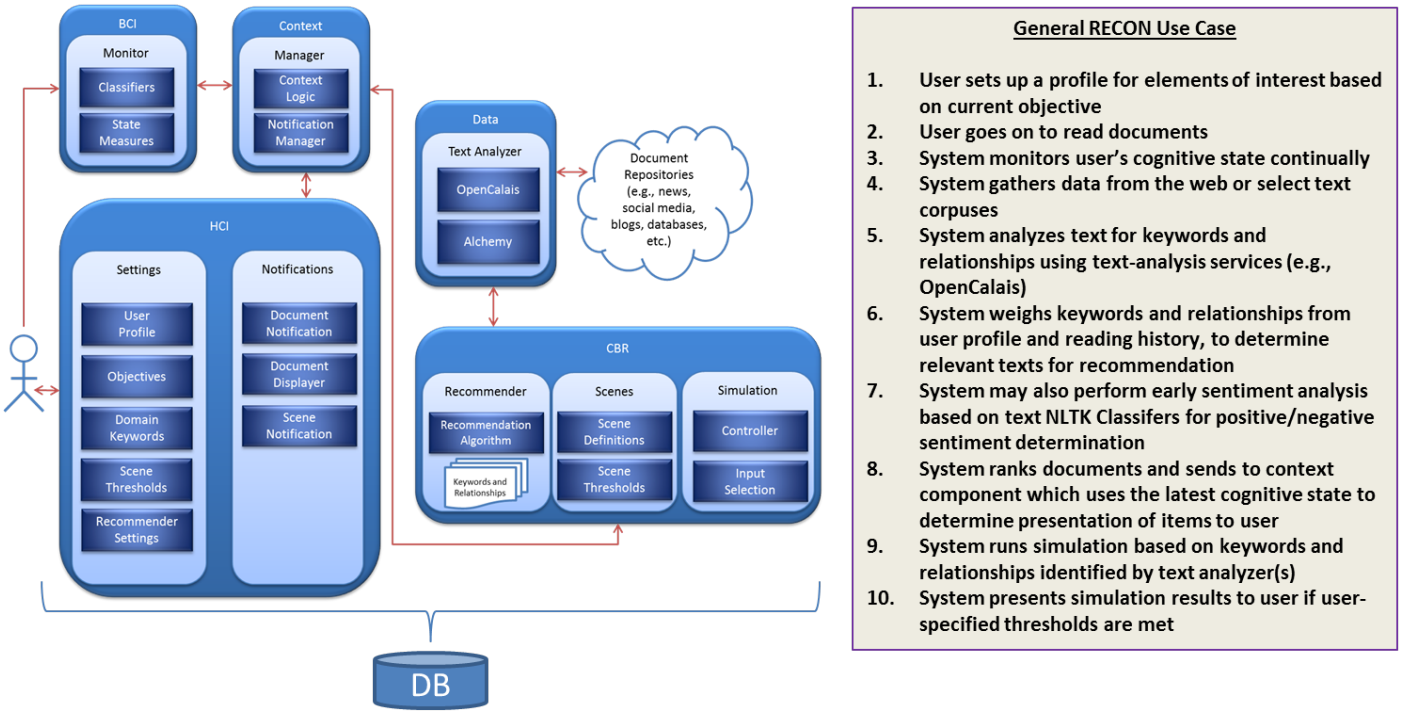


Figure 3. The RECON architecture in more detail: RECON determines context and makes recommendations based on user preferences, brain-state data, and simulation results in order to help alleviate the problem of information overload. The primary use-case is also shown.

B. *RECON Architecture*

The detailed RECON architecture, outlined previously in [8], is shown in Figure 3. It incorporates the following five layers, which are described below and which correspond directly to the use case.

1) Brain-Computer Interface (BCI) Layer: The BCI layer of RECON is concerned with monitoring and classifying the psycho-physiological state of the user. Its actions include monitoring electroencephalography (EEG) signals (from the user's headset) and classifying the user's implicit contextual state based on established models of EEG analysis (e.g., measures based on excitement, relaxation, alertness, and stress levels) [16].

In general, BCIs provide mechanisms to acquire, transform, and classify bio-signals into learned states that may then be applied as factors in determining system actions. There are multiple sources of bio-signals related to brain activity that could be useful in real-time, such as EEG, functional Near Infrared Spectroscopy (fNIRS), and functional magnetic-resonance imaging (fMRI); although, for practical purposes, wearability and wireless communication become important criteria for selecting a BCI paradigm. Ideally, the acquisition technology should be as unobtrusive to the user as possible, and not restrict mobility, while simultaneously obtaining brain signals and transmitting these to a processing module. As such, only fNIRS and EEG approaches are relevant solutions, as other techniques involve large and expensive units (see [17], [18] for a discussion on these techniques).

In this work, EEG approaches have been selected as they have the added benefit of being readily available in the form of commercial headsets, including two candidate wireless headsets: the Emotiv EPOC and the Neurosky Mindwave [19], [20]. Whereas the Neurosky Mindwave provides only a single sensor, the Emotiv EPOC has been selected as it provides brain-signal data from multiple sensors, at sites relevant for estimating the states useful for the analyst scenario. In particular, these states include arousal and valence, as in [17], [19], [21]. *Arousal* represents a measure of activation versus inactivation (i.e., being ready to act or not), while *valence* represents a measure of pleasure versus displeasure (i.e., attraction or withdrawal). These have been selected as early measures and can be swapped for new measures, such as alertness and load, as identified in the literature [19], [22]. The combination of arousal and valence provides a circumplex of affect, or emotion, as discussed in [23], which in RECON allows the system to deduce whether analysts are in states such as alert, happy, content, bored, or angry and thereby determine an appropriate response action.

The process whereby signals are translated from raw EEG into state classifications involves the following phases: i) acquisition of raw EEG; ii) pre-processing and noise reduction using discrete wavelet transforms; iii) EEG feature extraction, according to known formulae based on work such as [19], [22]; and iv) classification of features into states, using a classifier trained on labelled datasets (such as [24]) to output levels of arousal and valence.

2) Human-Computer Interface (HCI) Layer: The HCI layer is concerned with monitoring and managing the RECON interface. Its main actions include identifying the current task of the analyst (e.g., whether the analyst is logged into the system, currently setting objectives, or reading a document) and adapting the graphical user interface (GUI) (e.g., whether specific portions of the display should be hidden so as to min-

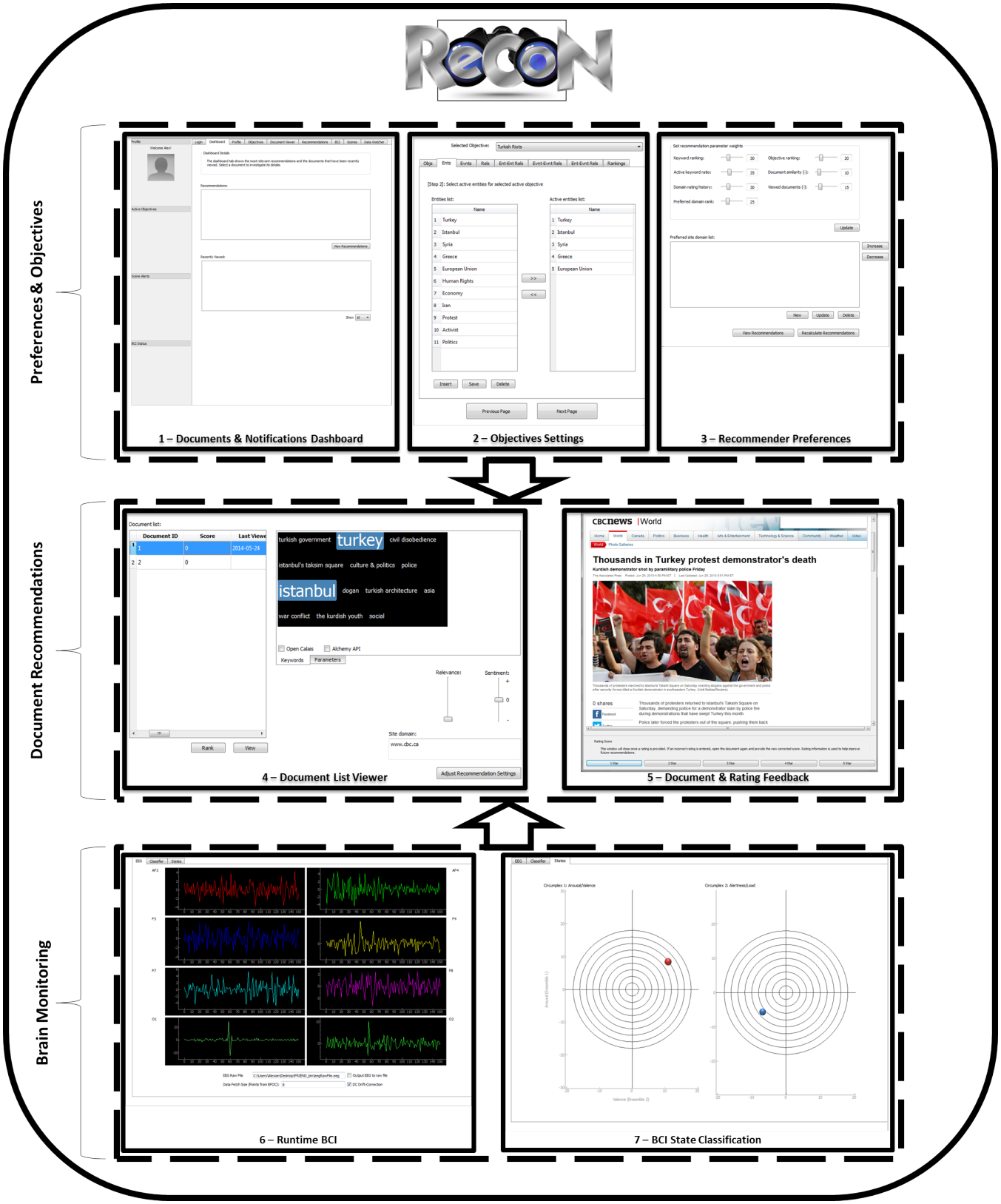


Figure 4. The main RECON interfaces.

imize analyst distraction). This layer also provides the analyst with a means to read and rank documents, set objectives and preferences, and interact with other RECON components, as outlined in the use case from Figure 3.

The main RECON interfaces are shown in Figure 4. These include GUI elements for the following: i) ranked document lists and notifications dashboard, which allows the analyst to view the highest-ranking documents (for both individual and team recommendations) and any system notifications; ii) objectives settings, which allow the analyst to specify and rank his/her target objective(s) in terms of entity and event keywords and relationships; iii) recommender preferences, which allow the analyst to fine-tune the recommendation parameters, such as the analyst's preferred website domains; iv) document viewing and v) ratings feedback, which allow the analyst to view all document recommendations for active objectives, browse document metadata, open documents for viewing, and provide a relevance rating for each document the analyst opens for reading; and vi) BCI runtime and vii) states, which allow the analyst to view the EEG inputs and the corresponding output states. In addition, support also exists for the following GUI elements (omitted from the figure for space reasons): user and team profile, which allows the analyst to set profile information and manage team members; and BCI setup, which allows the analyst (or an administrator) to specify the active BCI classifiers and start or stop real-time EEG monitoring. Together, these interfaces represent the visual and interactive components of RECON.

3) Data Layer: The data layer is responsible for collecting and monitoring incoming data (e.g., documents), analyzing them, and storing the results in a database for later use. This layer monitors preset bodies of textual data, in the form of documents, from websites and other corpuses. When new documents arrive, they are processed using existing text analysis engines, such as AlchemyAPI [25] and OpenCalais [26], in order to provide tagging of documents (e.g., keyword metadata) and sentiment markers. This approach is modular, and specific text-analysis engines, with different underlying ontologies, may be substituted based on the specific domain needs of the active situation(s)-of-interest.

4) Case-Based Recommender (CBR) Layer: The CBR layer is concerned with ranking processed data (i.e., tagged documents) based on specific recommendation criteria so as to present the analyst with a recommendation of the most relevant system data available. The actions include storing the analyst-specific recommendation criteria (e.g., relevant keywords, preferred website domains, and user rating history) and updating the recommendation list based on newly processed documents and user action and feedback (e.g., which documents have been read and what ratings were provided by the analyst). It also provides an algorithm for document recommendation based on keywords and relationships, the facility for defining scenes and monitoring scene thresholds, and a simulation controller and input-selection mechanism for managing simulations. These simulations are used as scene conditions to model aspects of the situation-of-interest (e.g., system dynamics can be used to estimate particular threshold variables over time, governed by causal-loop diagrams that incorporate variables and their interrelationships, linked to specific objective keywords [27]).

The algorithm developed for the recommender, depicted in Figure 5, makes use of three separate subroutines to calculate the latest recommendation score for all documents in the system pertaining to a specified date range and user. The *getRecommendations* function sets the analyst-specified date range and queries the database checking for i) updated or new analyst-specified objectives (case a) and ii) new documents (case b). If any changes have been made to the objectives (i.e., case a), such as the addition of a new keyword, the *gatherDocumentProperties* function is applied to all documents in the date range; however, if no objectives have been added or changed but new documents are found (i.e., case b), the function is applied only to the new documents. If neither case occurs, the *calculateRecommendationScore* function is called directly. This prevents the system from having to recalculate fixed document properties each time the recommendation algorithm is used. The *gatherDocumentProperties* function retrieves the relevant documents (based on the applicable case), and, for each document, gathers the fixed document properties and compares the document keywords to the keywords and relationships associated with the current active objective(s) of the analyst. The *calculateRecommendationScore* function is then applied. This function retrieves the document properties stored by the previous function and applies analyst-specified modifiers as weights that impact the resulting recommendation score for each document. Such a multi-step mechanism allows the latest recommendations to be computed, taking into account the analyst's most recent ratings and preference settings, without needing to recompute fixed document properties each time a recommendation update is requested.

Specifically, for each document-objective pairing, a document score, σ_{doc}, is calculated according to the following equation:

$$\sigma_{doc} = \phi_{key} * \pi_{key} + \phi_{rel} * \pi_{rel} + \phi_{site} * \pi_{site} \quad (1)$$

where ϕ_{key} represents the ratio of keywords in the document compared to the total number of keywords specified in the objective; ϕ_{rel} represents the sum of the relevance of matching keywords as ranked in the objective; if the document comes from a site domain that is preferred by the user, ϕ_{site} represents a positive value based on the relative rank position of the preferred site domain within a user-specified list (zero otherwise); and the π_* values represent the preference weighting of each factor (a real number between zero and one inclusively) as specified by the user.

This score is then used in the determination of the document's resulting recommendation score, σ_{rec}, calculated according to the following equation:

$$\sigma_{rec} = \sigma_{doc} + \phi_{obj} * \pi_{obj} - \phi_{sim} * \pi_{sim} - \phi_{acc} * \pi_{acc} \quad (2)$$

where σ_{doc} represents the fixed document properties according to the equation above; ϕ_{obj} represents a positive value based on the relative ranking of the active objective compared to all active objectives; ϕ_{sim} represents how similar the current document is compared to all documents viewed already by the user in reference to this objective; ϕ_{acc} represents whether or not the document has already been accessed (i.e., viewed) by the user for this objective: one if true and zero if false; finally, the π_* values represent once again the preference weighting of each factor as specified by the user. By reducing the score based on document similarity, the aim is to increase document *coverage* [28] using the following heuristic: recommend to the user the highest scoring documents that have the least similarity compared to documents already viewed.

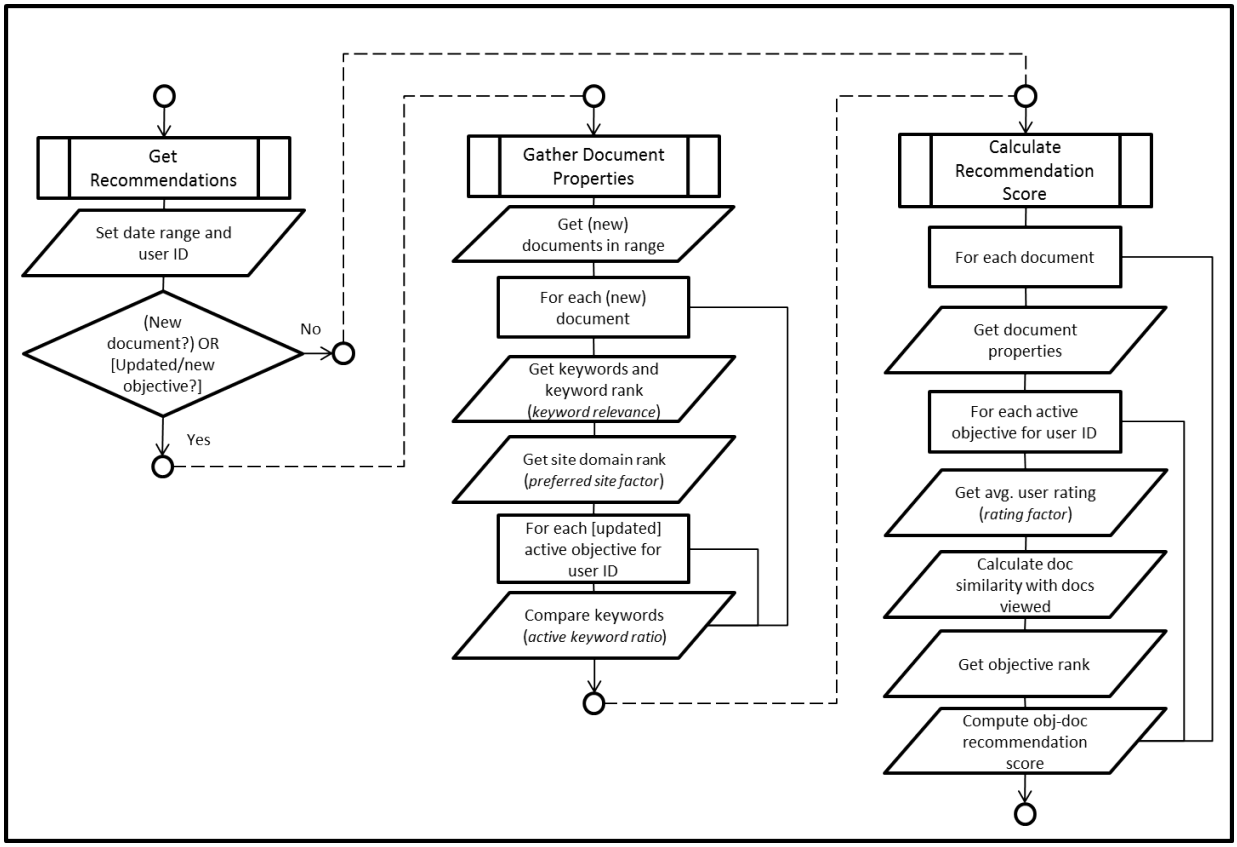


Figure 5. The algorithm used to make document recommendations in RECON.

Moreover, the CBR component is supported by the scene and simulation components, which, together, manage scene notifications coming from the system to the analyst. The scene component is concerned with the creation and monitoring of scenes. The actions of this component include storing scenes created by the analyst, monitoring incoming processed data to determine if specific scene conditions have been met, and issuing a notification if a scene's condition threshold has been reached. The simulation component, on the other hand, is concerned with the creation and execution of simulations, whose results act as particular scene conditions. The actions of this component include storing the location of external simulation models or the models of internal simulations supported directly by the component, as well as the input parameters that are passed to these simulations. This component supports a combination of simulation paradigms (e.g., system dynamics, discrete-event, and multi-agent simulation) to better match the representational requirements of the current situation (e.g., system-level or individual-level concerns) with the most appropriate paradigm [27]. Other actions include executing the simulations and storing the results in the system database.

5) Context Layer: The context layer is the final layer and is concerned with assessing the overall current context of the analyst. The actions of this layer include acquiring all available implicit and explicit context from the other layers, determining the current context of the analyst, managing what information is sent to the analyst (e.g., from the recommender layer), and initiating available GUI interventions (via the HCI layer) to reduce experienced information overload on the part of the analyst. It is from these other layers that the context layer collates and makes sense of this information.

In addition to detecting the user's contextual states, it is important to be able to operationalize this information to improve adaptive system behaviour. Consequently, the context layer makes use of a well-known cognitive model for the intelligence analysis domain as part of its adaptation strategy. The COntextual COntrol Model (COCOM), described in [29] and based off the work of Hollnagel [30], is a foundational model outlining four different control states that can be in effect for an analyst based on the amount of time remaining to make a decision. These states—strategic control, tactical control, opportunistic control, and scrambled control—represent a continuum from strategic control, where the decision-maker has sufficient time to plan, to scrambled control, where the decision-maker is faced with very limited (to potentially no time) to plan. These control states, when applied to the sense-making loop in Figure 1, result in a set of parameters that can be used in determining the analyst's cognitive mode in light of an unfolding event.

As shown in Figure 6, the COCOM model has been fitted to support RECON's context-management approach. In RECON, two classes of recommendation exist: (i) documents, which consist mainly of new input from text sources; and (ii) scenes, which can include things such as newly simulated situation projections. Together, these represent the two "cases" that are recommended by the system according to the current state of the analyst. While documents tend to provide information on a particular situation-of-interest that is more specific in nature

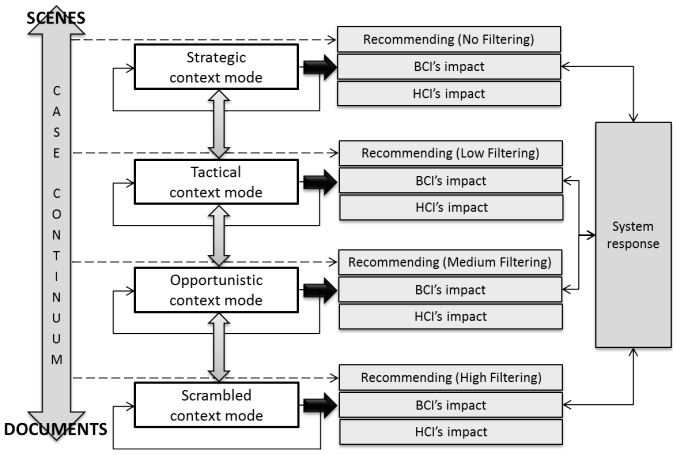


Figure 6. The COCOM model applied to RECON context management (adapted from [29], [30]).

and may arrive at any time, scenes tend to reflect higher-level, strategic and tactical outlooks that would generally be created only when the user has sufficient time. However, alerts related to these scenes can come at any time (independent of the user's context mode), in the same way as document recommendations do.

An analyst can be in one of four context modes, determined by the context layer using both explicit and implicit contextual sources. These context modes, ordered according to decreasing time-available-to-plan and directly based on a mapping from COCOM control states, are as follows: the *strategic context mode*, where there is a significant amount of time remaining before a decision is required; the *tactical context mode*, where there is sufficient time remaining to consider alternate avenues; the *opportunistic context mode*, where time is limited; and the *scrambled context mode*, where time is very limited (or has run out) and a decision must be made as soon as possible. Moreover, these four modes map to system adaptation of recommendations and alerts. When the strategic context mode has been identified, the system performs no special filtering of document recommendations and scene alerts, allowing the analyst to view a wide range of information and scene-projections, some of which may not be "on-task." Likewise, when the tactical context mode is deduced by the context layer, the system makes use of low filtering, whereby more off-task alerts and scene projections are not directly presented to the analyst. In the opportunistic context mode, the system uses a medium level of filtering for recommendations and alerts, allowing only near-task and on-task information to be shown to the analyst. Lastly, in the scrambled contextual mode, the system adapts with high filtering of incoming recommendations and alerts, presenting only on-task information to the analyst. The determination of on-task recommendations and alerts is based to a large extent on an analyst's preferences, such as the ranking of current objectives, keywords, and scenes, while the determination of the current context mode, as discussed earlier, is based on a combination of analyst-context data from both explicit and implicit sources.

This section has presented the RECON use case and system in detail, and it is envisioned that such a unique combination of layers, enhanced through the use of explicit and implicit context management, can better support analysts in performing their tasks by satisfying the different information "rights" mentioned in Section II, thereby improving the machine's ability to effectively assist the analyst and reduce cognitive overload. RECON furthers the goals of alleviating human-cognitive overload in two ways. First, it does so by developing a system capable of sensing and classifying the user's contextual state, including brain state using a brain-computer interface. Secondly, it does so by adapting to the user's context and recommending relevant information to the user based on the system's level of context-awareness. While the vision and use case for RECON have been presented and a proof-of-concept system implemented, they remain to be validated experimentally. However, the foundation for each component is empirically supported by recent literature. In particular, for the brain-computer interface, work such as [19], [31], [32] demonstrates that real-time brain-state classification is indeed viable. In terms of human-computer interfaces, work such as [33], [34] highlights the benefits of integrating HCI and BCI for adaptive systems. Lastly, in terms of both context and recommendation, studies such as [35], [36] underscore the effectiveness of context-based recommendation. These research foundations enable a merger of technologies as presented in RECON, and such a merger requires new architectural design considerations in order to achieve a more cohesive software system. These considerations are examined in the following section.

IV. Architectural Design Considerations

This section outlines eight key architectural design considerations, relevant to the design of adaptive systems at large. These have been grouped into three categories—human modelling, human-machine interaction, and human-machine synergy—according to the taxonomy shown in Figure 7. This taxonomy is described below, followed by a critical discussion of the eight considerations.

A. Considerations Taxonomy

The *human-modelling* category of the taxonomy, shown in Figure 7, relates to design considerations affecting how the human is modelled within the computer system. As a key component of human-machine systems, human modelling acts as the mechanism used by the machine to better understand and represent the user. Two relevant considerations are considered in the next subsection: *model selection*, relating to how user mental states are determined; and *model calibration*, relating to how these specific models are initialized and tuned.

The *human-machine interaction* category refers to the design considerations involving human-machine interaction. These relate to the interface between the human and the machine, with a particular emphasis on the human-in-the-loop acting as a critical component of the overall system [37]. Three considerations are examined in the subsequent subsection: *model transparency*, relating to the extent to which the user understands (and needs to understand) the internal mechanisms driving the system; *user feedback*, relating to how the machine system receives feedback from the user; and *contextual inputs*, relating to how the system receives or collects contextual input from the user (i.e., implicitly or explicitly).

Lastly, the *human-machine synergy* category deals with those considerations affecting the effectiveness of the human-machine team in accomplishing the overall goal of the system.

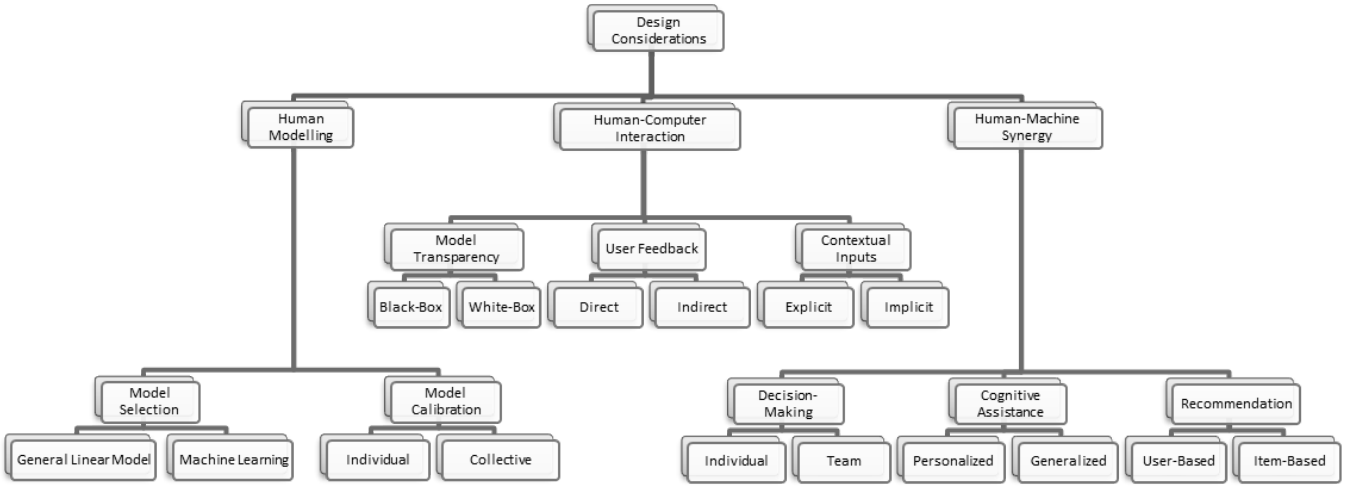


Figure 7. Taxonomy of architectural design considerations for context-aware systems.

Three particular considerations are examined in the following subsection: *decision-making*, relating to the support for organization-wide decision-making within the system; *cognitive assistance*, relating to the the extent of personalization supported within the system; and *recommendation*, relating to the mechanism(s) used in suggesting different information items to the user.

B. The Eight Considerations

The eight architectural design considerations are discussed in detail below and are presented in the order in which they appear in the taxonomy, left to right.

1) Model Selection: General Linear Model vs. Machine Learning: The first consideration, relating to modelling the user's brain state, is concerned with the selection of an appropriate method for capturing the underlying pattern of cerebral activity associated with a given state, namely statistical analyses based on the General Linear Model (GLM) or Machine Learning (ML) algorithms. The GLM approach has a proven track record among the neuroscience community and has robust analysis software available [38]. However, complex non-linear relations cannot be "discovered" using this method (i.e., the underfitting problem) due to the linearity constraint, yet this contraint makes the GLM very robust to noise (e.g., measurement error and intrusions from confounding factors), thus minimising the overfitting problem [39]. Underfitting occurs when the model lacks sufficient functional flexibility to capture a phenonemon, while overfitting occurs when the model's flexibility allows it to "fit" both the true regularities in the data as well as false, noisy patterns, leading to an overestimation of the model's accuracy [40]. ML algorithms, on the other hand, including those related to data mining, provide highly flexible models capable of discovering complex patterns in datasets. However, this flexibility raises a potential vulnerability to overfitting, which must be considered.

2) Individual Calibration vs. Collective Calibration: The second consideration involves model calibration, which can occur at either the collective level or at the individual level. The former results in a single model for all potential users, while the latter results in a distinct, customized model for each user. Individual modelling has the disadvantage of requiring additional overhead for calibration, including a separate data collection for each user in order to extract an individualized model. Nevertheless, this approach may be essential for attaining high levels of model accuracy, particularly in cases when the average of the collective is the result of idiosyncratic patterns [41], [42]. Alternatively, individual differences can be treated as noise, although this could potentially lead to underfitting of the user state.

3) Model Transparency: White-Box vs. Black-Box: The third consideration, related primarily to human-machine interaction, is how much of a model's inputs, logic, and resulting assessment to display to the user. A transparent, "white-box" model may increase user trust in the system, but there is also the risk of fostering mistrust in situations where the user disagrees with the model or does not understand it. Furthermore, a significantly complex display could adversely impact the understandability of the model to the user. On the other hand, a "black-box" approach, in which model details are completely hidden from the user, may also foster doubt and mistrust in the system. As such, this consideration relates to the classic invisibility dilemma: balancing between minimizing distractions from the primary task of the user and providing added value through explicit interaction with the model [43].

4) Direct Feedback vs. Indirect Feedback: The fourth consideration involves whether or not to collect feedback directly from the user in order to improve the underlying models used in adaptation. Direct feedback incorporates the user in the learning process by requiring a manual response about the performance of the system, which in turn guides the fine-tuning of model parameters. This feedback assumes a level of expertise on the part of the user. Moreover, the frequency of direct feedback must be considered, as it may unnecessarily burden the user if required too often [44]. Conversely, indirect feedback allows the system to acquire the necessary inputs for the fine-tuning parameters without involving the user directly. This has the benefit of allowing the user to remain on task, while simultaneously allowing the system to improve its adaptation, as often as needed. There is a tradeoff, however, in terms of the accuracy of the learning mechanism, as some

aspects may not lend themselves readily to being deduced indirectly.

5) Explicit vs. Implicit Contextual Inputs: The fifth consideration involves knowledge about user context, which is central to system adaptation. This context evolves according to events and changes occurring during system operation either by explicit interactions from the user (e.g., a user manually indicates current context parameters such as time pressure) or implicit interactions based on the situational context (e.g., automatic data monitoring and sensor-based classification). Explicit context affords the user a sense of control over the system and provides contextual data that may not be otherwise available. However, a system that relies too heavily on this type of context may add to the workload of the user, in terms of providing a larger amount of information manually to the system, and may require a more complex graphical-user interface and additional tasks that may interfere with the user's ability to focus on the task-at-hand. Conversely, a system that emphasizes implicit context frees the user from tedious data input operations, but requires the system to automatically monitor data and perform reasoning to infer the user's contextual information. This demands a significant a priori development effort for effective user-state and contextual classification models.

6) Individual vs. Team Decision-Making: The sixth consideration relates to the manner in which decision-making is performed, and how this can be assisted through technology. In some organizations, individual contribution is valued more than the collective, if not explicitly then implicitly through their reward structure; however, in dealing with complex systems, a broad range of expertise should be drawn upon [45], [46]. Moreover, if people are tired or overloaded, being able to offload a particular task to a more alert member of the team can help the organization make more effective use of its resources [46]. In fact, having the ability to promote a networked culture is seen as a vital step in addressing complex issues [45]. This is not so much a technical issue, as it is an organization-design issue. However, technology can be brought to bear to facilitate or promote the spread of this culture, and such considerations form an integral aspect of system design [47].

7) Personalized vs. Generalized Cognitive Assistance: The seventh consideration, dealing with the human factor, relates to the individual needs of a user. Not everyone is the same, and people have different cognitive abilities and assistance needs. Sometimes people may remember a lot of information at once and be able to recall it; other times they may wish to offload some of this information to a machine. Moreover, each person may view the situation from a different perspective, or have different sub-problems to address. As such, a certain level of user customization may be desirable. However, tradeoffs must be considered. For example, the cost of such customization may be seen as being too high at times. This can be from the point-of-view of the system designer, as it provides more freedom to the user and less predictability on the part of the system [47], but also on the part of the user, who may not see or understand the value in customization. If customization is desired, a possible solution to the latter problem is to provide tutorials and walkthroughs to help guide users in better understanding the benefits of customization.

8) User-Based vs. Item-Based Recommendation: The eighth consideration involves the mechanism used to rank recommendation items. Traditionally, there are two main approaches to recommendation. The first, known as collaborative filtering, involves creating a user profile and comparing it to the profiles of other users. The objective is to find a subset of closest neighbours, whose preferred items can then be recommended to the user under consideration. This is good for situations in which having "trusted friends" can be of benefit to the recommendation (e.g., when wanting to be given a recommendation for a book or movie); however, it does suffer from the cold start problem in which establishing an accurate profile for a new user takes time and many rating samples. The second approach, known as item-based filtering, uses the properties of the items themselves. The objective here is to find similar items to those items the present user ranked most highly. The benefits of this approach are that newer items have just as much chance of being selected as older items and it is good when the set of other system users is small. However, items must be comparable, so it might not work as effectively for recommendations involving a wide-range of differing items (e.g., Amazon). Hybrid methods have also been proposed [28] in which combinations of different recommendation algorithms are used in tandem. These have the effect of mitigating the weaknesses of any one approach, and different techniques may be more suited to specific domains.

This section introduced eight important architectural design considerations for supporting context-awareness in human-machine systems. This together with the previous section, which detailed the RECON proof-of-concept software system, are the focus of the following section. Specificially, it examines how the proposed key architectural considerations have been applied in RECON.

V. Applying the Architectural Design Considerations to RECON

In this section, the application of the eight architectural design considerations presented in Section IV is described according to the five layers of RECON presented in Section III: namely, brain-computer interface (BCI), human-computer interface (HCI), data, case-based recommender (CBR), and context layers. The strengths and benefits of the RECON architecture, resulting from the conscientious application of these considerations, are also discussed, along with possible improvements.

A. Brain-Computer Interface (BCI) Layer

The following design considerations apply to the BCI layer and are presented according to the taxonomy in Figure 7:

- *Model Selection:* A machine learning approach has been selected to recognize dynamic, non-stationary EEG signals. In particular, the use of a neural-network (neuro-fuzzy)-based classifier approach provides a method for making sense of brain EEG data, which is well-supported in literature and provides a generalizable approach to classification, allowing additional state measures to be incorporated [48], [49].
- *Model Calibration:* All selected models are initially calibrated using a collective approach with labelled, pre-existing datasets for training the classifier. This means that there is no need for a lengthy training

session by the analyst prior to using the system. However, there is potential to adapt the classifier to individual characteristics based on performance feedback provided to the system over time.

- *Model Transparency:* A black-box approach to classification has been selected for pattern-recognition, which means that specific details are hidden from the analyst. However, EEG inputs, feature configurations, and state measures can be inspected. This hides low-level model details, which should only be modified by an expert, while providing an overview of the BCI process, which may facilitate the user's trust in the model's output.
- *User Feedback:* An indirect feedback strategy has been selected for the BCI layer, as direct user feedback is obtained elsewhere in RECON and can be applied to the BCI in terms of performance-based adaptation. The benefit of this approach is that the analyst need not have BCI-specific expertise to update the BCI model; instead, this can be abstracted to other parts of the system.
- *Contextual Inputs:* Implicit contextual inputs, in the form of EEG signals, have been selected to deduce an analyst's cognitive state. This allows the system to unobtrusively monitor the user, without requiring explicit input by the user with regards to their current mental state. This is important as the user might not be aware of their own current psycho-physiological state.

B. Human-Computer Interface (HCI) Layer

The following considerations apply to the HCI layer:

- *Contextual Inputs:* Both explicit and implicit contextual inputs have been used in this layer. In particular, explicit inputs have been selected to allow users to set current objectives and preferences, which are used for recommendation and filtering. This has the benefit of allowing the system to know exactly what the analyst is trying to accomplish. Implicit context is also gathered by the system to determine the current task of the user (e.g., whether the user is logged into the system, currently setting objectives, or reading a document), which can be used to adapt system notification levels.
- *Decision-Making:* Functionality for both individual and team decision-making has been supported within the HCI layer. This provides the facility for analyst teams to share objectives and recommendation results and manage team membership, thereby supporting shared situational awareness.
- *Cognitive Assistance:* Functionality for personalized cognitive assistance has been provided in the form of adaptable document lists and notification areas within the HCI layer. This has the benefit of streamlining the presentation of content based on the analyst's current cognitive state and specified preferences (e.g., how the ranking should occur).

C. Data Layer

The following consideration applies to the data layer:

- *Model Transparency:* A black-box approach has been selected for this layer. This is because existing solutions, which are readily (and even freely) available, themselves are black-box solutions, and the development of new approaches to text analysis is outside the scope of the current project. While the specific mechanisms driving a text-analysis engine may not be relevant to an analyst, the ontologies are. As such, within RECON, the results from different text-analysis engines can be activated by the user to compare which is performing best for a particular objective.

D. Case-Based Recommender (CBR) Layer

The following considerations apply to the CBR layer:

- *Model Transparency:* A black-box approach has been selected for recommendation. This has the benefit of hiding the low-level implementation details of the algorithm from the analyst. However, the analyst is still enabled to provide inputs in the form of objectives and tuning preferences that influence the recommender. While the algorithm has been implemented, further testing is required, which may necessitate modifications in order to achieve recommendations that are both relevant and highly diverse when compared to documents the analyst has already seen (as per [28]).
- *User Feedback:* Direct user feedback has been selected for the recommendation layer. The analyst is required to rate each recommended document he/she reads in terms of its relevance to the associated active objective. This has the benefit of allowing the system to obtain immediate and targeted feedback, without unduely burdening the user. This feedback is used to improve future recommendations and can also be used to provide indirect performance feedback to the BCI layer (e.g., correlating BCI state measures with the most-recently rated document to determine a relevance measure for the analyst).
- *Contextual Inputs:* Explicit inputs have been selected for specifying the current situational context of the analyst. This involves setting and defining the current active objective(s), which include the specification of entity and event keywords and relationships between these keywords, along with the relative ranking, used for prioritization, of the objectives and objective components. RECON guides the analyst in defining his/her objectives, and recommender effectiveness is determined precisely by how well the recommendations align with these objectives. The benefit of explicit, guided contextual input is that the system obtains the problem situation directly from the analyst, without having to deduce such potentially-complex and diverse context implicitly. Moreover, this allows for targeted recommendations that can later be refined by modifying the user-specified context.
- *Decision-Making:* Both individual and team decision-making have been selected for the recommendation layer. This means that an analyst can view recommendations related to both his/her objectives, as well as those shared by the analyst's team members. Together, these have the benefit of supporting increased

situational awareness across the team or organization, while still effectively catering to the particular interests or needs of individual analysts. The currently implemented approach could be improved by allowing further refinements in terms of what exactly is shared to other team members (e.g., only share recommendations above a certain rating threshold); however, the team recommendations list can be sorted according to the recommendation rating, among other properties, allowing the analyst to quickly filter items in the list.

- *Cognitive Assistance:* Personalized cognitive assistance has been selected for this layer. This is in the form of scenes. A scene can be defined by the analyst in order to represent a particular aspect of the problem situation he or she wishes to offload to the system. The benefit of this approach is that the analyst is free to define (or not) as many scenes as may prove beneficial. This degree of customization allows expert users to capitalize on personalized cognitive assistance, which can include keyword tracking and multi-paradigm simulations for what-if analysis.
- *Recommendation:* Item-based recommendation has been selected for this layer, where the "items" in this case refer to documents. This approach uses the features of the document (such as tagged keywords), rather than properties of other users who may have read the document, in order to determine the document's relevance to the current analyst. In the traditional approach to item-based recommendation, the properties of other items the analyst has viewed and ranked would be used in the relevance calculation [28]. However, in RECON, because an analyst specifies precisely what he or she is interested in at the present time through objectives, this explicit context is used instead. The benefit is that the analyst's most-recent intentions are always incorporated into the RECON recommendations, rather than using potentially outdated intention history, as is the case in the traditional approach.

E. Context Layer

The following considerations apply to the context layer:

- *Model Transparency:* A white-box approach has been selected for the context layer. The COCOM model has been adapted for context management, and the application of this well-known method allows analysts to better understand the system behaviour (e.g., in terms of its filtering actions). This has the benefit of promoting confidence in the system's behaviour, while also allowing for potential future improvements involving direct feedback from the user with regards to the context mode determination of the system (e.g., the analyst feels he or she is in the tactical context mode, while the system has determined that the current mode is scrambled context).
- *Contextual Inputs:* Both implicit and explicit contextual inputs have been selected for this layer. Implicit context is obtained from the BCI component as well as from the HCI activity log, which shows the current action the analyst is performing in the system. Explicit context is obtained primarily through the analyst's definition of objectives. The benefit of a combined approach is that context that can be deduced by the system (e.g., current psycho-physiological state) is acquired without direct involvement of the analyst, while context that is more difficult to ascertain automatically (e.g., changing objectives and priorities) can be acquired explicitly from the analyst. This promotes an effective balance between explicit analyst involvement in the context adaptation process and system usefulness, allowing the analyst more time to focus on important tasks such as sense-making.
- *Decision-Making:* Currently, individual decision-making has been selected for this layer. This comes in the form of contextual mode classification at the analyst level, which determines how recommendations and alerts are filtered to the individual. A future improvement would be to support team decision-making at this layer. This would take the form of recommendations and alerts being filtered across a team of analysts, where a particular notification would be sent to the analyst best-suited to receive it (e.g., an analyst who is not determined to be overloaded and for whom the content of the document is meaningful, i.e., it matches with at least one of the analyst's individual or team objectives). This would have the benefit of sending the right information to the right person at the right time, three key criteria of the five "rights" discussed in Section II.
- *Cognitive Assistance:* A generalized cognitive assistance approach has been selected for the context layer. This comes in the form of the COCOM-based model, which defines four possible contextual modes an analyst may be in, as well as the actions the system performs in response to an analyst being in a particular state. This has the benefit of providing individuals with adaptive responses, while not requiring direct feedback from the analyst in order to do so (e.g., an analyst specifying how much filtering to perform for a particular contextual mode). However, as a possible future improvement, personalized fine-tuning could be incorporated into the system, but would require additional testing to determine the trade-off of added personalization.

These architectural design considerations are inter-woven into the fabric of the resulting system architecture, which speaks to their interconnectedness. As such, it is important to explore these considerations carefully when designing RECON-like systems, as a change in one location can easily impact other parts of the system. Figure 8 summarizes the design considerations discussed in this section, organized according to the five layers of the RECON architecture and the considerations presented in Section IV. To underscore the uniqueness of what is being proposed in this paper, the following section examines related work concerning architectural design considerations.

VI. Related Work

The preceding sections have identified architectural design considerations for adaptive context-aware systems and their application in the recent RECON implementation for the

		Architectural Design Consideration							
		Human Modelling		*Human-Machine Interaction*			*Human-Machine Synergy*		
		Model Selection [GLM vs. Machine Learning]	*Model Calibration [Individual vs. Collective]*	*Model Transparency [Black-Box vs. White-Box]*	*User Feedback [Direct vs. Indirect]*	*Contextual Inputs [Explicit vs. Implicit]*	*Decision-Making [Individual vs. Team]*	*Cognitive Assistance [Personalized vs. Generalized]*	*Recommendation [User-Based vs. Item-Based]*
RECON Component	*Brain-Computer Interface*	Machine Learning	Collective	Black-Box	Indirect	Implicit	N/A	N/A	N/A
	Human-Computer Interface	N/A	N/A	N/A	N/A	Explicit and Implicit	Individual and Team	Personalized	N/A
	Data	N/A	N/A	Black-Box	N/A	N/A	N/A	N/A	N/A
	Case-Based Recommender	N/A	N/A	Black-Box	Direct	Explicit	Individual and Team	Personalized	Item-Based
	Context	N/A	N/A	White-Box	N/A	Explicit and Implicit	Individual and Team	Generalized	N/A

Figure 8. Design considerations applied to RECON context management.

intelligence analysis domain. These considerations have been motivated by known HCI dilemmas and the cognitive overload problem faced by analysts [1]. The literature on context-aware systems is vast, as seen in [14], and architectures have been proposed that are similar to RECON.

For example, in [50], the authors propose a multi-module approach for a context-aware system middleware having the following modules: a reasoning engine, learning engine, context predictor, access controller, and context integrator. Likewise, in [51], the authors propose an ontology-based decision-support system for the military domain, comprising multiple agents responsible for decision support, user information, available sensors, information services, and context management. Even though these are similar in that they combine multiple tiers for context management, these approaches do not share the same layers as RECON, nor is their emphasis on reducing information overload. While much attention in the literature has focused on such architectures, relatively little has been devoted to the design considerations guiding the development of these systems [14], which is a core focus of this paper. In this section, related work on architectural design considerations is presented and compared with those relevant to RECON, as have been presented in Section IV.

In [47], twelve HCI dilemmas are discussed in the context of supervisory control. A significant number of these are philosophical in nature, such as who should ultimately be in *control*, the human or the machine, and what is the "right" balance between automation and control. These considerations do not relate directly to the RECON system. However, others are more application-oriented, such as the role that *trust* plays in the system and how much trust should be placed in the results coming from automation. Another is how much *free will and creativity* to allow on the part of the user versus having a system that is completely predictable from the designer's perspective. These two are directly applicable to RECON in terms of both model transparency (i.e., trust) and cognitive assistance (i.e., the extent of user-involvement in the personalization process).

In [52], four design considerations are proposed that directly support two distinct, but related aspects: i) intelligibility of system behaviour and ii) accountability of human users. These considerations include informing the user about the current *capabilities and understanding* of the contextual system, which is in-line with the role of trust in [47] and the idea of model transparency as proposed in Section IV. System *feedback* is also a key feature outlined in [52] and is meant to inform the user about both the consequences of a particular action prior to its being enacted (feedforward) and notification about what the user has done following the action (confirmation). To this end, the authors propose that *identity and action disclosure* be incorporated into a system as part of an audit trail. This differs from RECON in that system feedback is defined by the user through scenes, which then allows the system to provide alerts that are objective-focused. Lastly, *control* is also emphasized and it is noted that the user should have the ultimate control over any actions he or she may be held accountable for. However, in RECON, because the user is restricted to a limited set of actions, including setting objectives and rating documents, this type of control is not a major consideration.

In [53], the design considerations presented are concerned with providing maximum flexibility to business processes. Key issues include how business processes can be *conceptualized* and applied to process models in general. This is not a major concern for RECON, which has been designed and implemented to support the established intelligence analysis cycle outlined in Section II. Another consideration presented in [53] relates to the *contextual variables* used to capture and assist with the business processes. The authors speak to the relevance and the observability of these contextual variables (e.g., some variables might not be observable and may need to be inputted by the user). In RECON, this notion is related to the balance between explicit and implicit contextual inputs wherein objectives and scenes are set explicitly by the analyst, while user-state classification results implicitly from the BCI assessment. The final issue mentioned in [53] is how business processes can be supported in the face of changes to context.

This *flexibility* also relates to the contextual-input consideration in RECON as recommendations, which support the intelligence analysis business process, automatically adapt to changes in context defined by the analyst through explicit objective and scene definitions.

Other researchers have focused on more singular considerations. For example, in [54], the researcher's major design considerations revolve around *enterprise collaboration* and how trust can be improved to support decision making across the entire enterprise. This effort speaks to collaborative management systems and the relevance of research focusing on networked businesses (or "holons" [46]). The major artifact stemming from this work is a table of trust criteria that can be used when implementing such systems. This relates closely to RECON's consideration of individual versus team decision-making, which is crucial as organizations increasingly must coordinate efforts in order to manage complex situations. Finally, in [55], the major focus area is *privacy* and how it should be managed. The authors propose providing the user with full control over which applications should be given information about the user's present location. While an explicit design consideration was not mentioned, the design considerations implicitly revolved around the problem of privacy and how best to ensure it. In terms of RECON, privacy is not a key consideration, as sharing information is central to sensemaking among analysts, and those receiving an analyst's information are considered to be trustworthy. It remains to be fully investigated how much implicit information, like an analyst's brain-state classification from EEG signals, is appropriate to be shared with other members of the organization. Such a policy would necessarily need to be determined on an organizational basis.

Each of the foregoing, while being related to context-aware systems, presents a unique perspective that highlights distinct architectural design considerations. As seems natural, these considerations are heavily motivated by the problem under investigation. For example, in [55] the authors focus on privacy, so the architectural considerations in the paper relate to how best to support user-controlled privacy. For RECON, the uniqueness of the solution, in terms of combining many different, yet relevant and supporting techniques, brings with it a unique set of considerations, which are not always considered in one system. Hence, the architectural design considerations described in Section IV offer a foundation from which future research attempting to create an adaptive, context-aware solution to the problem of cognitive overload in intelligence analysis can begin.

VII. Conclusion and Future Work

With its focus on cognitive offloading and high-relevance system recommendations, RECON targets the adaptive context-aware systems domain for intelligence analysts through a unique five-layer architecture having an explicit human-factors view of context management. Eight key architectural design considerations have been proposed herein for context-aware support, and their application to the implemented RECON system has been presented. Moreover, previous architectural discussions have been extended in this paper with a detailed recommendation algorithm and a cognitive model for context classification.

It is expected that in situations involving information overload, uncertainty, and time pressure, the effectiveness of intelligence analysts can be significantly improved through context-aware adaptive systems, where these design considerations have been conscientiously applied. However, there remains room for more comparisons and discussion of these considerations in light of future system implementations. Also, more practical testing of the architectural implementation is required to ascertain its ability to support analysts. As part of future work, a human-in-the-loop experiment will investigate the effectiveness of the RECON implementation and approach in reducing cognitive overload based on the principles of adaptive-context management.

Acknowledgment

This work was funded by Defence R&D Canada, by Thales Research and Technology Canada, and by a research partnership grant from the Department of National Defence of Canada and the Natural Sciences and Engineering Research Council of Canada. The authors would also like to acknowledge the reviewers, both from this journal and the ADAPTIVE 2014 conference, for their thoughtful feedback.

References

[1] D. Lafond, R. Proulx, A. Morris, W. Ross, A. Bergeron-Guyard, and M. Ulieru, "HCI dilemmas for context-aware support in intelligence analysis," in ADAPTIVE 2014, The Sixth International Conference on Adaptive and Self-Adaptive Systems and Applications, 2014, pp. 68–72.

[2] G. Baxter and I. Sommerville, "Socio-technical systems: From design methods to systems engineering," Interacting with Computers, vol. 23, no. 1, 2011, pp. 4–17.

[3] K. Vicente, The Human Factor: Revolutionizing the Way People Live with Technology. Routledge, 2004.

[4] Central Intelligence Agency, "The work of a nation," Library of Congress, Tech. Rep., 2009.

[5] P. Pirolli and S. Card, "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," in Proceedings of International Conference on Intelligence Analysis, vol. 5. Mitre McLean, VA, 2005, pp. 1–6.

[6] ——, "Information foraging," Psychological review, vol. 106, no. 4, 1999, p. 643.

[7] D. M. Russell, M. J. Stefik, P. Pirolli, and S. K. Card, "The cost structure of sensemaking," in Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems. ACM, 1993, pp. 269–276.

[8] W. Ross, A. Morris, M. Ulieru, and A. B. Guyard, "RECON: An adaptive human-machine system for supporting intelligence analysis," in Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on. IEEE, 2013, pp. 782–787.

[9] E. S. Patterson, D. D. Woods, D. Tinapple, E. M. Roth, J. Finley, G. G. Kuperman, and H. E. Directorate, "Aiding the intelligence analyst in situations of data overload: From problem definition to design concept exploration," Institute for Ergonomics/Cognitive Systems Engineering Laboratory Report, ERGO-CSEL, 2001.

[10] M. Chesbro, "Intel-cyclopedia: A guide to sources of information for the intelligence community," Homeland Security Digital Library, 2011, retrieved: April 2014.

[11] Ö. Yılmaz and R. C. Erdur, "iConAwa – An intelligent context-aware system," Expert Systems with Applications, vol. 39, no. 3, 2012, pp. 2907–2918.

[12] A. Schmidt, M. Beigl, and H.-W. Gellersen, "There is more to context than location," Computers & Graphics, vol. 23, no. 6, 1999, pp. 893–901.

[13] G. Fischer, "Context-aware systems: The 'right' information, at the 'right' time, in the 'right' place, in the 'right' way, to the 'right' person," in Proceedings of the International Working Conference on Advanced Visual Interfaces. ACM, 2012, pp. 287–294.

[14] J.-y. Hong, E.-h. Suh, and S.-J. Kim, "Context-aware systems: A literature review and classification," Expert Systems with Applications, vol. 36, no. 4, 2009, pp. 8509–8522.

[15] D. Gouin, V. Lavigne, and A. Bergeron-Guyard, "Human-computer interaction with an intelligence virtual analyst," in Proceedings of Knowledge Systems for Coalition Operations, IHMC, Pensacola, FL, 2012, pp. 1–5.

[16] A. Morris and M. Ulieru, "FRIENDs: Brain-monitoring agents for adaptive socio-technical systems," Multiagent and Grid Systems, vol. 8, no. 4, 2012, pp. 329–347.

[17] B. Graimann, B. Allison, and G. Pfurtscheller, "Brain–computer interfaces: A gentle introduction," in Brain-Computer Interfaces. Springer, 2010, pp. 1–27.

[18] ——, Brain-Computer Interfaces: Revolutionizing Human-Computer Interaction. Springer, 2010.

[19] R. Ramirez and Z. Vamvakousis, "Detecting emotion from EEG signals using the emotive epoc device," in Brain Informatics. Springer, 2012, pp. 175–184.

[20] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review," Sensors, vol. 12, no. 2, 2012, pp. 1211–1279.

[21] A. Holm, K. Lukander, J. Korpela, M. Sallinen, and K. M. Müller, "Estimating brain load from the EEG," The Scientific World Journal, vol. 9, 2009, pp. 639–651.

[22] D. O. Bos, "EEG-based emotion recognition," The Influence of Visual and Auditory Stimuli, 2006, pp. 1–17.

[23] J. A. Russell, "Core affect and the psychological construction of emotion." Psychological review, vol. 110, no. 1, 2003, p. 145.

[24] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis using physiological signals," Affective Computing, IEEE Transactions on, vol. 3, no. 1, 2012, pp. 18–31.

[25] AlchemyAPI Website, URL: http://www.alchemyapi.com (Last accessed: 2014.11.12).

[26] OpenCalais Website, URL: http://www.opencalais.com (Last accessed: 2014.11.12).

[27] W. Ross, M. Ulieru, and A. Gorod, "A multi-paradigm modeling and simulation approach for system of systems engineering: A case study," in IEEE 9th International System of Systems Engineering Conference, 2014, pp. 1–6.

[28] L. Candillier, M. Chevalier, D. Dudognon, and J. Mothe, "Multiple similarities for diversity in recommender systems," International Journal On Advances in Intelligent Systems, vol. 5, no. 3 and 4, 2012, pp. 234–246.

[29] B. F. Gore et al., "Human performance cognitive-behavioral modeling: A benefit for occupational safety," International Journal of Occupational Safety and Ergonomics, vol. 8, no. 3, 2002, pp. 339–351.

[30] E. Hollnagel, "Context, cognition, and control," in Co-operative Process Management, Y. Waern, Ed. London: Taylor & Francis, 1998, pp. 27–52.

[31] S. Makeig, C. Kothe, T. Mullen, N. Bigdely-Shamlo, Z. Zhang, and K. Kreutz-Delgado, "Evolving signal processing for brain–computer interfaces," Proceedings of the IEEE, vol. 100, no. Special Centennial Issue, 2012, pp. 1567–1584.

[32] C. Mühl, B. Allison, A. Nijholt, and G. Chanel, "A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges," Brain-Computer Interfaces, no. ahead-of-print, 2014, pp. 1–19.

[33] A. Lécuyer, L. George, and M. Marchal, "Toward adaptive VR simulators combining visual, haptic, and brain-computer interfaces," Computer Graphics and Applications, IEEE, vol. 33, no. 5, 2013, pp. 18–23.

[34] D. Tan and A. Nijholt, "Brain-computer interfaces and human-computer interaction," in Brain-Computer Interfaces. Springer, 2010, pp. 3–19.

[35] U. Panniello, A. Tuzhilin, and M. Gorgoglione, "Comparing context-aware recommender systems in terms of accuracy and diversity," User Modeling and User-Adapted Interaction, vol. 24, no. 1-2, 2014, pp. 35–65.

[36] P. G. Campos, I. Fernández-Tobías, I. Cantador, and F. Díez, "Context-aware movie recommendations: An empirical comparison of pre-filtering, post-filtering and contextual modeling approaches," in E-Commerce and Web Technologies. Springer, 2013, pp. 137–149.

[37] T. Tsiligkaridis, B. Sadler, and A. Hero, "A collaborative 20 questions model for target search with human-machine interaction," in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, May 2013, pp. 6516–6520.

[38] G. D. Hutcheson and N. Sofroniou, The multivariate social scientist: Introductory statistics using generalized linear models. Sage, 1999.

[39] M. A. Pitt, W. Kim, and I. J. Myung, "Flexibility versus generalizability in model selection," Psychonomic Bulletin & Review, vol. 10, no. 1, 2003, pp. 29–44.

[40] S. Roberts and H. Pashler, "How persuasive is a good fit? a comment on theory testing." Psychological review, vol. 107, no. 2, 2000, pp. 358–367.

[41] W. Estes and W. T. Maddox, "Risks of drawing inferences about cognitive processes from model fits to individual versus average performance," Psychonomic Bulletin & Review, vol. 12, no. 3, 2005, pp. 403–408.

[42] P. N. Mohr and I. E. Nagel, "Variability in brain activity as an individual difference measure in neuroscience?" The Journal of Neuroscience, vol. 30, no. 23, 2010, pp. 7755–7757.

[43] A. Schmidt, M. Kranz, and P. Holleis, "Interacting with the ubiquitous computer: towards embedding interaction," in Proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies. ACM, 2005, pp. 147–152.

[44] S. Schiaffino and A. Amandi, "User–interface agent interaction: personalization issues," International Journal of Human-Computer Studies, vol. 60, no. 1, 2004, pp. 129 – 148.

[45] D. S. Alberts and R. E. Hayes, "Power to the edge: Command... control... in the information age," DTIC Document, Tech. Rep., 2003.

[46] W. Ross, A. Morris, and M. Ulieru, "NEXUS: A synergistic human-service ecosystems approach," in Sixth IEEE Int'l Conf. Self-Adaptive and Self-Organizing Systems Workshops (SASOW), 2012, pp. 175–180.

[47] T. B. Sheridan, "HCI in supervisory control: Twelve dilemmas," in Human error and system design and management. Springer, 2000, pp. 1–12.

[48] A. Subasi, "EEG signal classification using wavelet feature extraction and a mixture of expert model," Expert Systems with Applications, vol. 32, no. 4, 2007, pp. 1084–1093.

[49] ——, "Application of adaptive neuro-fuzzy inference system for epileptic seizure detection using wavelet feature extraction," Computers in Biology and Medicine, vol. 37, no. 2, 2007, pp. 227–244.

[50] W. Chun-dong, L. Xiao-qin, and W. Huai-bin, "A framework of intelligent agent based middleware for context aware computing," in Natural Computation, 2009. ICNC'09. Fifth International Conference on, vol. 6. IEEE, 2009, pp. 107–110.

[51] S. Song, K. Ryu, and M. Kim, "Ontology-based decision support for military information systems," in Applications and Technology Conference (LISAT), Long Island Systems. IEEE, 2010, pp. 1–5.

[52] V. Bellotti and K. Edwards, "Intelligibility and accountability: human considerations in context-aware systems," Human–Computer Interaction, vol. 16, no. 2-4, 2001, pp. 193–212.

[53] M. Rosemann and J. C. Recker, "Context-aware process design: Exploring the extrinsic drivers for process flexibility," in The 18th International Conference on Advanced Information Systems Engineering. Proceedings of Workshops and Doctoral Consortium. Namur University Press, 2006, pp. 149–158.

[54] P. Kaur, S. Ruohomaa, and L. Kutvonen, "Enabling user involvement in trust decision making for inter-enterprise collaborations," International Journal on Advances in Intelligent Systems, vol. 5, no. 3 and 4, 2012, pp. 533–552.

[55] F. Dorfmeister, S. Feld, and C. Linnhoff-Popien, "ALPACA: A decentralized, privacy-centric and context-aware framework for the dissemination of context information," International Journal On Advances in Intelligent Systems, vol. 7, no. 1 and 2, 2014, pp. 223–236.

Enhancing Robustness through Mechanical Cognitivization

Gideon Avigad*, Wei Li†, and Avi Weiss*
*Department of Mechanical Engineering
ORT Braude College of Engineering, Karmiel, Israel
Email: {gideona, avi}@braude.ac.il
†Department of Automatic Control and Systems Engineering
University of Sheffield, Sheffield, UK
Email: wei.li11@sheffield.ac.uk

***Abstract*—The common approach for training robots is to expose them to different environmental scenarios, training their controllers to have the best possible commands when untrained scenarios are encountered. When humans train they do the same. They try new manipulations by performing within different environments. However, humans training (and in fact development from infancy to maturity) also includes a type of training which, although claimed to improve cognitive capabilities, has not, to date, been adopted for the training of robots. This type of training involves the restriction of manipulation capabilities while performing different tasks, e.g., climbing with just one hand. Recently a research that facilitates functions instead of a mechanical systems that aims at exploring the invigorating idea that such training, would enhance the robustness of robots, has been published. This type of training has been termed as Mechanical Cognitivization. In the current paper, the preliminary published results are detailed and more elaborated examples are given. Specifically, it is shown that the Mechanical Cognitivization based training improves the performances when performing within untrained environments and when malfunctions occur. The advantages of the suggested training are highlighted through facilitating a comparison between two schemes that include a common neural net (with no training of restricted modes) and the recently introduced Mechanical Cognitivization based neural net for which the training includes training of restricted modes. The results highlight the advantages of Mechanical Cognitivization based training in enhancing robustness.**

***Keywords*—Cognitive robotics, developmental robotics, evolutionary algorithms**

I. Introduction

The use of robots in performing industry-related tasks and operating within hazardous environments has become ubiquitous. Yet robots are rarely used in everyday tasks that are performed mainly by humans. Indeed, humans exhibit truly amazing competencies in performing arduous and complicated tasks that may involve changing working conditions (scenarios), while controlling and maneuvering their multi-degrees-of-freedom body. By repeatedly executing different tasks, the human brain learns how to control the complex human body. Two human activities are of interest to the current research. The first is associated with training for participation in sports and athletic activities. For example, when training, climbers often use different techniques such as climbing with one hand tied behind the back, climbing sloping walls with no hands, or climbing blindfolded. Clearly, such actual climbing conditions are not expected. Rather, these are all training techniques intended to improve climbers' sense of balance and movement skills. Restriction of movement as a training method is found in other sports as well, among them swimming (e.g., swimming with just one hand or without using the legs) and the martial arts (Fighting blindfolded). The second human training activity of interest here is also related to restricted movement and involves the way human capabilities develop from birth. The training of babies' minds begins on a body that is not yet fully developed. In contrast to new-born calves or horses, for example, human babies cannot stand, walk or run. Evolution has dictated a slow rate of development among humans and has forced the use of restricted capabilities. Could this be because in many situations, only some of the body's competencies are used so that the body must also be trained for these sub-manipulations? Note that many sports advocate starting young in order to let the body and mind adapt to the demands of the sport. In [1], we suggested exploring the novel idea of enhancing the robustness of robots by training them while taking into consideration both their final bodies/embodiments and their restricted modes (less capable versions). Such training has the potential to enhance the robustness of robots in performing untrained maneuvers as well as in coping with malfunctions and unexpected working conditions.

For elucidating the idea behind Mechanical Cognitivization (MC), suppose that a robotic climber (CR) needs to be developed, so that it is capable of climbing a wall with poles sticking out of it. The left panel of Figure 1 depicts one possible mechanical configuration (body) for such a CR. The CR now must be trained to maneuver up and grasp one of the poles (A or B). The idea suggested here is that during the training of the controller of this CR, not only should this body be utilized but also its restricted modes. Two such restricted modes are depicted in the middle and right panels of the figure. Clearly, performing such a maneuver by utilizing one of the restricted modes might be associated with degraded performances (e.g., larger integral of the square error, measured while considering the planned and actual maneuver performed). Restricted modes may include the following restrictions: a) using only some of the mechanical capabilities, such as preventing some of the

links from moving – that is, if the robot has four arms/links, it will be restricted to use only two or three of them or will be prevented from using its gripper; b) restricting the movement of the arms/links to less than their full possible extent; c) deliberately imposing friction at the joints; d) changing the stiffness of the links; or e) restricting the actuator performance, for example by reducing the power supply to the actuators or using weaker actuators (smaller motors).

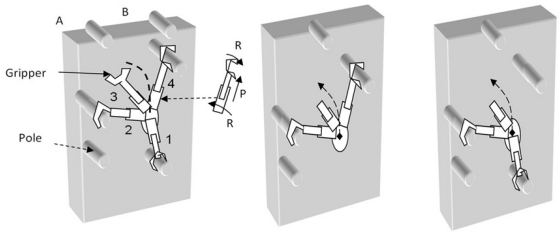


Fig. 1. CR having four links is trained using all of them (left panel) and using restricted modes (middle and right panels). This figure was from [1].

The work in [1] explains the idea, which is demonstrated by using functions. In the current paper, more elaborated examples are given. It is shown that the MC based training improves the performances when performing within untrained environments and when malfunctions occur. The paper is organized as follows. Section II discusses the needed background, which includes cognitive architectures and developmental robotics. The methodology is given in Section III. It includes the way to train and test the different MC related schemes. Next, in Section IV, the success of MC in enhancing robustness is demonstrated through using a basic example (in Subsection IV-A) and further elaboration (Subsection IV-B). A discussion and envisaged future work are presented in Section V.

II. Background

Over the past several decades, a great deal of research attention has been directed at cognition and its implementation for artificial brains. The inspiration provided by human beings toward producing a machine that will copy human abilities is evident. Different models of cognition have been adopted to produce artificial cognitive systems or cognitive architectures. Cognitive architectures [2] represent attempts to create unified theories of cognition, i.e., theories that cover a broad range of cognitive issues, among them attention, memory, problem-solving, decision-making and learning. These theories consider several aspects, including psychology, neuroscience, and computer science. Examples of such architectures are the Soar system [3],[4], and ACT-R [5]. Some of these architectures have been claimed to be more adequate than others for use as cognitive brains for robots. This distinction (see, e.g., [6]), is rooted in the differences between the "cognitivist" and the "emergent" philosophies of cognition. The philosophy of emergent cognition contends that the relationship between the cognitive architecture and the body it is controlling (e.g., robots) is essential to the development of cognition, which is not the case for the "emergent" philosophy. The current paper deals with the"emergent" philosophy, because the learning directly depends on the availability of models describing the controlled entity. An associated philosophy is embodied cognition [7],[8], which states that cognition can be influenced and biased by states of the body and that abstract cognitive states are grounded in states of the body. Among the architectures that facilitate this view is the biologically plausible brain-inspired neural-level cognitive architecture proposed by Shanahan [9], in which cognitive functions such as anticipation and planning are realized through internal simulation of interaction with the environment. Burghart et al. [10] proposed a hybrid cognitive architecture for a humanoid robot that is based on the interaction of parallel behaviour-based components and a long-term memory sub-system utilizing a variety of representational schemas, including object ontologies and geometric models, Hidden Markov Models, and kinematic models. For a comprehensive survey of many of the approaches to model cognition and the resulting cognitive architectures, see [6].

Several approaches have been proposed to improve the response of artificial entities to specific stimulations by circumventing complex cognitive architecture. For example, the computational model of perception and action for cognitive robots discussed in [11] embraces the view that there is a direct route from perception to action that may bypass cognition [12]. A related approach is morphological computing (see, e.g., [13],[14],[15]), in which the idea is to design the mechanical structure to respond directly to a stimulus. This response is a result of the special morphology (shape, materials inter-relation among parts) of the structure. For example, in [16] the special features of a hand (Yoki hand) partially built from flexible deformable materials enable it to easily grasp different objects with no need for controller feedback. This notion has gained a great deal of interest, and for the past several years workshops have been dedicated to considering different aspects of morphological computing, such as artificial skin and stretchable sensors, compliant actuators and mechanisms, and soft materials in robotics.

In contrast the proposed research focuses on the enhancement of cognition by considering the mechanical structure, as is the case in morphological computing. Here, however, the cognitive architecture is of vital importance, and the mechanical structure and its possible restricted modes (permutations of the final structure) are utilized for training the cognitive architecture. This means that the mechanical structure is the driving force for the enhancement of cognition/learning. This enhancement of cognition by facilitating the mechanical structure of the robot has been termed as Mechanical Cognitivization (MC) [1].

Most relevant to the current paper are studies conducted by Mark Lee's group at Aberystwyth, UK. Their research is related to Developmental Robotics [17]. According to this approach, which is rooted in the way babies develop, cognitive development is achieved through staged growth of cognition as the sensomotoric competencies are gradually and sequentially improved. In several publications [18],[19],[20] Lee's group introduced and developed what they term as 'constraint lifting'. At each stage, learning takes place with

certain constraints imposed on the sensomotoric system. At the next stage, some of these constraints are removed or 'lifted'. For example, learning hand-eye coordination in manipulating a robotic arm has been investigated. In that case, as leaning progressed, constraints imposed on moving parts of the robot (e.g., using the fingers) were 'lifted'.

The current paper and the MC idea involve several basic differences from the works such as [19]: a) In contrast to the sequential staged growth, MC may be enhanced simultaneously. b) In the proposed approach, constraining manipulations may take place any time along the robot's life time and c) In one of the hereby proposed schemes, the knowledge gained through training in a restricted mode, may be preserved separately, and utilized when needed. As discussed above, cognition involves among other issues, learning, anticipation, conceptualization etc. The current paper deals just with the learning phase. For this reason and for the sake of focusing the current study on proving the applicability and impact of MC on robustness, we chose to utilize architectures, which are merely neural nets as was done in [1].

Previous studies on embodied cognition concentrated on how to construct a sophisticated artificial cognitive architecture, by utilizing one embodiment of the controlled entity. In contract here, we elaborate on the MC idea, which is aimed at fully exploiting the capabilities of any controller (artificial brain), by facilitating the learning of its embodiment including its related restricted modes (less capable embodiments). Here we further elaborate on the results attained in [1], through considering different combinations of restricted modes, allowing a more in depth look into the suggested learning approach.

III. Methodology

In order to elucidate the MC idea and to demonstrate its potential, mathematical functions are used here as was done in [1], instead of a CR or any other mechanical system. Representing the "environment" (the climbing wall) to which the CR has to adapt to (able to climb in the best way), is a polynomial function $Y(x)$, of order m where x is a vector of inputs (e.g., location of poles). In other words $Y(x)$ may be viewed as a planned route for the robot to follow. The CR's controller is a neural net (NN) for which the outputs are coefficients of a polynomial of order n, $y(x) = a_1x^n + a_2x^{n-1} + ... + a_n$. Each output may be viewed as a control signal (here it is a coefficient) to a motor of a manipulator that moves a robotic arm. The sum of the arm's movements results (through kinematics) in the location of the CR on the wall. Here, this summing is represented by $y(x)$. The correlation between the CR case and the function representation, is summarized using Figure 2.

A. Training schemes

The training of the NN may be enhanced by, e.g., minimizing the square Error, $Error = (Y(x) - y(x))^2$ averaged over the available function's points. In order to train the net to give an output, which is adequate to the environment (the function of order n), the artificial learning system was set as depicted in Figure 3.

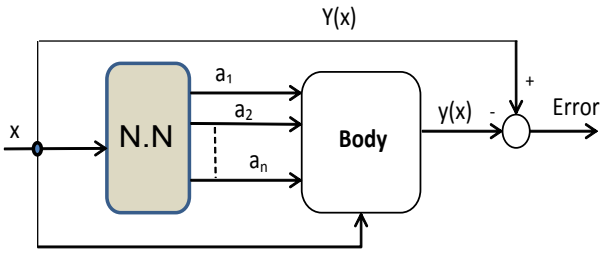


Fig. 2. The correlation between the CR and the related function representation.

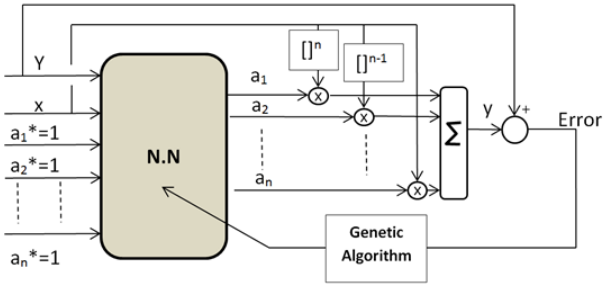


Fig. 3. The artificial NN with no restricted modes.

The input to the NN, is a list of k, x and corresponding $Y(x)$ values, which are fed sequentially to the net. The Net has extra n inputs (flags), namely:$A = [a_1^*, a_2^*, .., a_n^*]$. Each flag may be assigned a binary value. If $a_i^* = 1.0$, it means that the net's i-th output is not prohibited and the related coefficient (link or DOF) participates in evaluating $Y(x)$. If $a_i^* = 0.0$, the i-th coefficient would be disregarded and the net is trained to produce just $n-1$ outputs in order to still fit, in the best way, to the original function (environment), which is of order m.

Definition: If $\neg\exists a_i^* = 0$ then the net is training/operating in a non-restricted mode. If $\exists a_i^* = 0$ then the net is training/operating in a restricted mode.

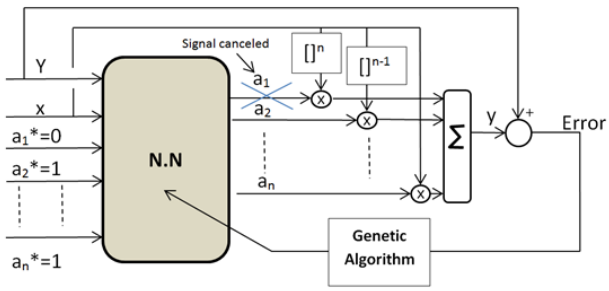


Fig. 4. The artificial NN training setting for a restricted mode.

While training, two NN schemes are considered. The first is merely the common NN, termed here as CNN, for which all flags' values are one. The other training involves a series of training sessions, which include a training in the non-restricted mode as well as a series of different restricted modes training. The same net is trained for accommodating all of

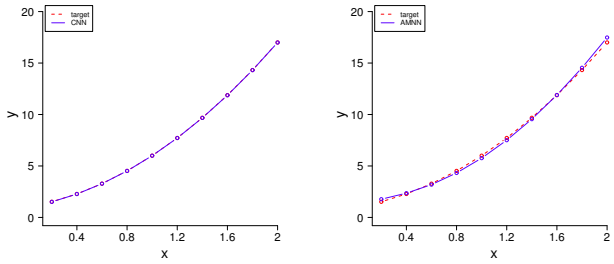
the restricted modes and therefore this trained net is termed here as Amalgamated Modes Neural Net (AMNN). In [1], for each restricted mode, the training of the AMNN exploited not more than the system's available resources (K pairs). K/n of the inputs, were pairs fed to the net together with all the flags set to one, as was in the non-restricted mode. For the next K/n of inputs, the output a_1 has been prohibited and the corresponding flag was set to zero, $a_1^* = 0$ and so forth. In the current paper this approach is not maintained. Here, the training at each mode, facilitates the entire set of available examples. The reasons for this change include: a) It is possible that the number of available training points is limited, leading to insufficient training resources for each restricted mode, b) It is conceivable that training the restricted modes and the non-restricted modes would require more resources. When human train, it is acknowledged that for getting better (e.g., by climbing with one hand) more training time and commitment is mandatory. The training of the different neural nets was done using the $(\mu + \lambda)$ evolution strategy with self-adaptive mutation strengths [21], where $\mu = 50$ and $\lambda = 50$.

IV. Testing the success of MC in enhancing robustness

Testing the different schemes for their robustness, is done here through considering two different uncertainties that involve untrained-for changes. These changes include, a) Malfunction: where one of the coefficients (robot's links/DOF) is prohibited. In such a case a more robust scheme would be the one for which the error with respect to the original function is smaller. This means that although malfunction occurs, the system is aiming at doing its "job" in the best way, and b) Environmental change: where the function is no more the original trained-for function (a new climbing route etc.). While testing the different schemes the following is assumed. The MC related neural net namely the AMNN has a feedback from the net's output (robot's links) such that if one or more malfunction, their related flags are changed to zero. In the case of the AMNN system, deliberately setting flags to zero for restricting movements along specific DOF, may be done. Decision on such a deliberate restriction may be done by a higher level controller that, e.g., uses vision to assess the accessibility to the target point. This means that a rational decision on, which of the modes to use may be done. When using just two fingers to lift a small object instead of using all fingers, humans are also using vision to make such a decision. For point b above another scenario may be envisaged. In such a scenario, manipulation takes place using the non-restricted mode. If the function is not satisfactorily estimated (target is not reached), another mode may be tried (the robot may retrieve to its initial configuration and retry). For the functions case, this means that the original function alters to a new one (new route) and a scheme that is more robust would be the one that adapts to the changed environment and acts to follow it with less error.

A. Example 1: Basic results

In this example an NN, which serves as the controller is to approximate a function of order two ($m = 2$), which was arbitrarily chosen to be: $Y(x) = 3x^2 + 2x + 1$. The approximating function has been chosen to be of order three (n=3): $y = a_1x^3 + a_2x^2 + a_3x + a_4$. It is noted that justifying this redundancy for the current case is rather hard and for now the importance of redundancy may be only borrowed from the correlation to the fact that human mechanics is redundant. The CNN's NN is a forward neural network with five inputs (two for the x and corresponding $y(x)$ and three for the flags a_1^*, a_2^* and a_3^*), four hidden neurons (tansig activation functions were used), and four output neurons (for a_1, a_2, a_3 and a_4). For finding the best net, one hundred evolutionary runs, involving each 5000 generations, were run for each training mode. The best net is chosen such that the average of the fitness function, over all training points, is the smallest. Hear $k = 10$ such that: $x = [0.2, 0.4, ..., 2.0]$. Figures 5(a), 5(b) depict the approximation of the target function by the CNN and AMNN schemes, respectively.



(a) CNN (b) AMNN

Fig. 5. This plot shows the performance of the a) CNN system and b) AMNN system with all flags set to be one.

It can be seen that the CNN approximation is somewhat better than that of the AMNN. This is not surprising because the the AMNN is trained using different restricted modes, some of which are not adequate for the function at hand (e.g., training when $a_2^* = 0.0$). The superiority of the CNN over the AMNN is further depicted in Figure 6, where the fitness value over 5000 generations of the evolutionary strategy run, is shown. As can be seen in the figure, the AMNN training needs more time to attain a reasonable good performance, while for the CNN, good performance is achieved rather quickly. Nevertheless, it will be shown that the merit of using the AMNN will be apparent when robustness to untrained scenarios would be tested.

1) Malfunction in example 1: The first test of robustness involves testing the robustness of the two schemes to malfunctions. For the current example a malfunction means prohibiting one of the DOF (one coefficient of $y(x)$ is set to zero). The performances of the two schemes are tested while both are to reach the training points. The performances of these schemes are compared using a box plot that are depicted in Figure 7. In all of the following box plots the line inside the box represents the median of the data. The edges of the box represent the lower and the upper quartiles (25-th and 75-th percentiles) of the data, whereas the whiskers represent the lowest and the highest data points that are within 1.5 times the inter-quartile range from the lower and the upper

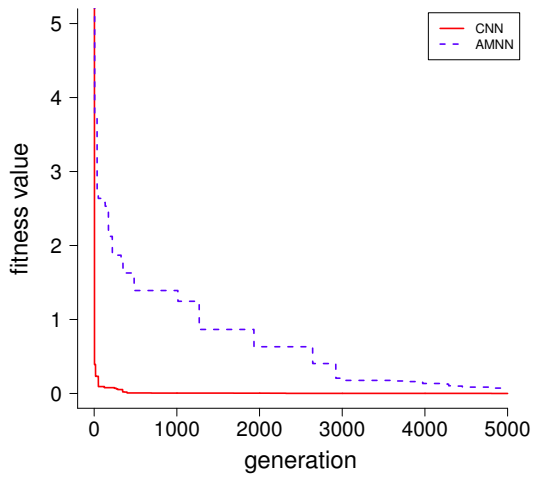


Fig. 6. The fitness of the best trained individual in each generation for the CNN and AMNN schemes.

quartiles, respectively. Circles represent outliers. In Figure 7, each box depicts the statistical data attained by one of the schemes over all training points, where the performances are the squared errors computed for each data point by: $Error(x) = (Y(x))^2 - (y(x))^2$.

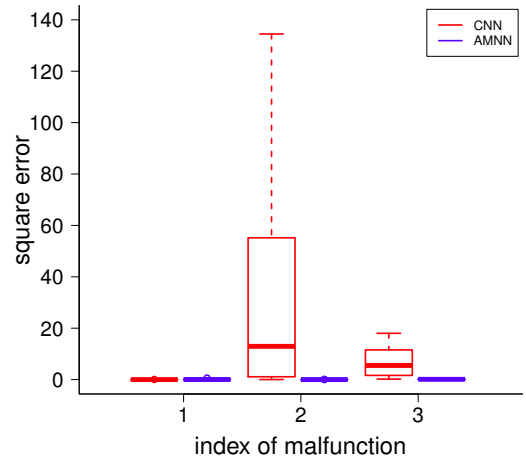


Fig. 7. The performance of CNN and AMNN system when malfunctions occur (index 1/2/3 corresponds to set the coefficients of third/second/first order of the function to zero). The p values for this figure (for the second and third DOF) are 0.0115 and 4.33e-05, respectively.

Restricting the first DOF (index "1" in Figure 7), namely a_1, does not result in any superiority of one scheme over the others. This is not surprising due to this DOF being of higher order than needed. However, when prohibiting the other two DOF, a_2 and a_3, which are designated by "2" and "3", respectively, the enhanced robustness of the AMNN, is highlighted. This is especially profound when a_2 is prohibited. The AMNN system performs significantly better than CNN system in the second and third case of malfunction (two-sided MannWhitney test, 5 percent significance level). Clearly, this advantage is built on training on restricted modes and preparing the scheme for such cases. Nevertheless, this training is just part of the training and the non-restricted mode is also part of the training.

2) Environmental Changes in example 1: For examining the robustness of the two schemes to environmental changes, the same x points as those of the training points are used, however, $Y(x)$ does not stay the same. For a mechanical system this would mean for the input, the output changes due to unexpected influences such as friction. To simulate such unexpected changes, the original function $Y(x)$ has been altered by changing its powers. The statistical data of changing the power two (second order) from one until three (skipping the original power, two) in steps of 0.2 and changing the power one (first order) from zero until two (skipping the original power, one), is depicted in Figure 8. Each data point in that figure is computed by computing the squared error between that function value at that point and the point reached by the scheme, averaged over all x points. The training followed the approach explained in Section IV.

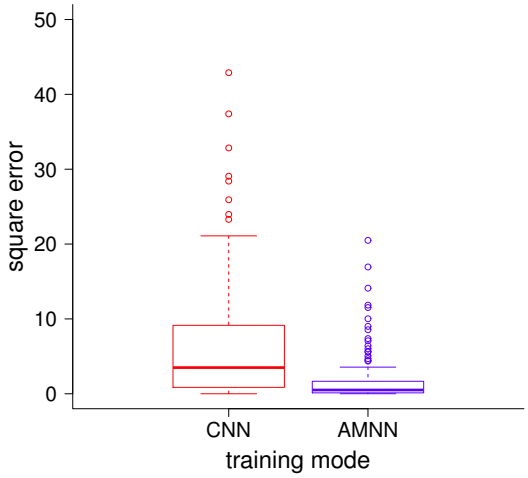


Fig. 8. The average fitness value of the CNN and AMNN system when changing the powers of the original function. The p value for this figure is 2.605e-11.

Depicting the results that are represented in Figure 8, it is clear that the MC related idea, which is realized through the AMNN scheme, promotes the robustness of the system when dealing with environmental changes as they are presented here (assuming the correlation between environmental changes and function changes).

For further testing the robustness to environmental changes, the original function is subjected to different changes, this time, instead of altering the powers, the coefficients are changed. In this case, the coefficients corresponding to the second order are changed from two to four in steps of 0.2, and the coefficients corresponding to the first order are

changed from one to three. There are total of $11 * 11 = 121$ combinations. The combination of three and two for the second and first coefficients, respectively, which corresponds to the original function, is excluded from the statistics (thus just 120 combinations are presented). The statistical data is presented in Figure 9. It can be easily seen that here again the AMNN performed significantly better than the CNN (two-sided MannWhitney test, 5 percent significance level).

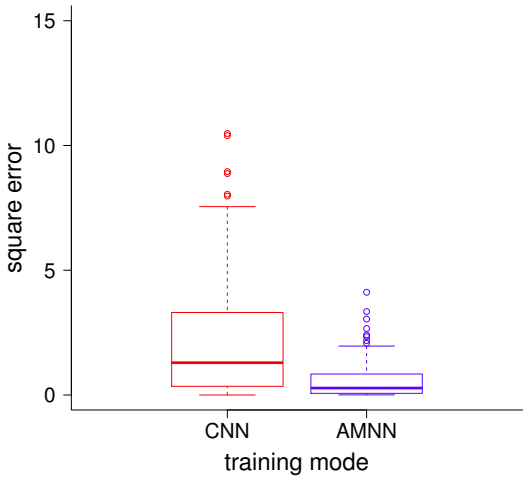


Fig. 9. The average fitness value of the CNN and AMNN system when changing the coefficients of the original function. The p value for this figure is 5.011e-11.

For elucidating the enhanced robustness of the AMNN scheme, Figure 10 depicts the performances of the AMNN scheme, designated in green and the CNN scheme, designated by blue and the changed environment related function shown by the doted red curve. Here the original (target) function has been altered to $Y = 3x^{1.5} + 2x^{0.5} + 1$.

The improved robustness of the AMMN is highlighted through its adaptation to the newly introduced function. Figure 11 depicts the same but for the case for which new coefficients are represented into the original function, namely, the function is altered to be $Y = 3.5x^2 + 2.5x^1 + 1$.

Again the competency of the AMNN scheme to respond to the needed changes, which is rooted in its ability to choose and activate/prohibit DOF, is highlighted.

B. Example 2: Elaborated results

In this section, the basic results, which were presented in Section IV-A, are elaborated and the approach is further tested. The main investigation concerns the effect of what restricted modes are utilized for the training phase. Consideration is given to the original CNN scheme and to three different training settings used for the AMNN scheme. The settings differ one from the other by the restricted modes that are used for the training. In all of these training settings, the non-restricted mode serves as one training mode and the other modes are as follows. AMNN1 is trained by a setting that is similar to the setting used for training the AMNN of

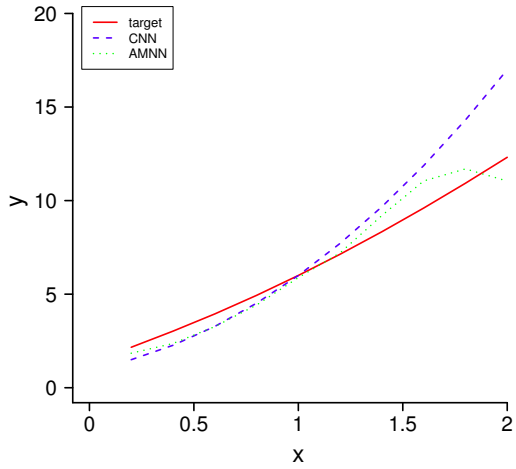


Fig. 10. One example showing how the CNN and AMNN system approximate a new function with different powers (simulating environmental changes).

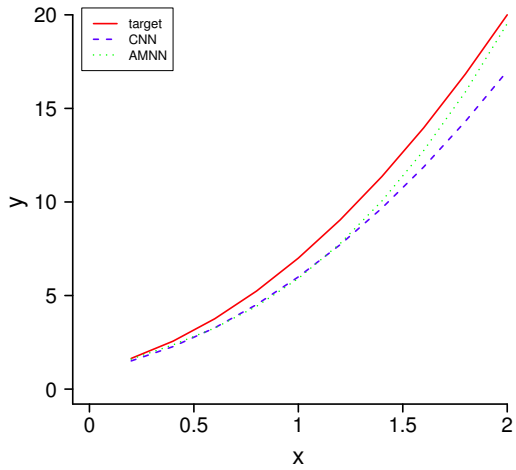
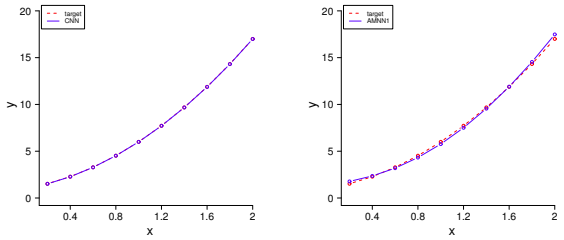


Fig. 11. One example showing how the CNN and AMNN system approximate a new function with different coefficients (simulating environmental changes).

Section IV-A. This means that three restricted modes are used ($k = 3$). In each of the restricted modes, one coefficient in y is cancelled. For AMNN2, each training involves restricting two modes. Because three coefficients are involved (a_1, a_2 and a_3, there are three restricted modes that are trained (again $k = 3$). For the AMNN3 setting the trainings used for both the AMNN1 and the AMNN2 are amalgamated. This means that restricting both one coefficient and two coefficients is practised, resulting in six restricted modes training ($k = 6$).

1) Normal Situation: Figure 12 shows the output (blue curves) of the best individuals (with the lowest fitness value) in the 5000^{th} generation for the CNN and AMNNs training in normal situation (In normal situation, all DOF of the robot are enabled, that is, all flags are set to 1.0.). Clearly, all train-

ing schemes approximate the target function in a reasonable accuracy. The fitness values of the the CNN scheme and the three AMNN's settings (AMNN1, AMNN2 and AMNN3) are $6.0 \cdot 10^{-5}$, 0.052, 0.125, and 0.072, respectively. The lower the fitness is, the better is the training. Therefore, the CNN training performs best in the normal situation. The advantage of the CNN scheme over all of the AMNN settings may be further highlighted by depicting Figure 13. The figure shows the fitness value of the best individuals in each generation during the evolutionary process for the CNN and AMNNs settings. As may be depicted, the fitness of CNN is always the lowest among the schemes in each generation. The training time required for the CNN is also much shorter than that of the AMNN settings. Comparing within the three AMNN settings, it may be observed that the fewer the DOF (i.e., k is smaller) is restricted during the training, the better average performance is attained.



(a) CNN (b) AMNN1

(c) AMNN2 (d) AMNN3

Fig. 12. This plot shows the output (blue curves) of the best individuals in the 5000^{th} generation for the CNN and AMNNs training to approximate the target function. The red (dotted) curve represents the target function.

In the following, the robustness of the different schemes and settings are examined based on the two unexpected changes, which were discussed in Section IV, namely malfunction and environmental changes.

2) Malfunction in example 2: For testing the robustness of the different schemes and related settings to malfunctions, each of them will be tested for different malfunction combinations. This means that the performances will be evaluated by restricting one coefficient at a time as well as when prohibiting more than one coefficient. Figure 14(a) shows the performance of each scheme when malfunctions occur using box plot. When the third DOF malfunctions (a_1 is prohibited), all the training modes can still approximate the target function well, since this DOF is redundant. For all the other cases of malfunction, the CNN scheme does not perform well, as may be comprehended from the relatively large square error

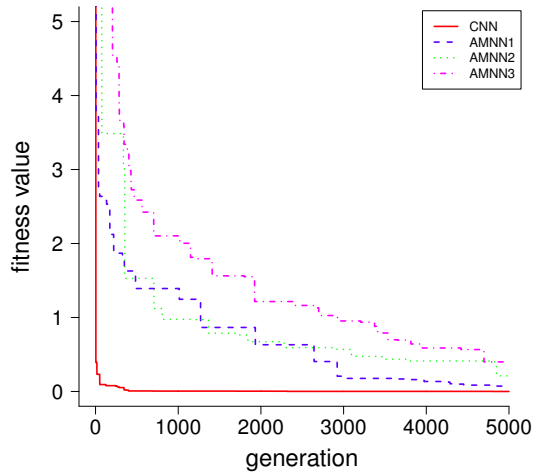
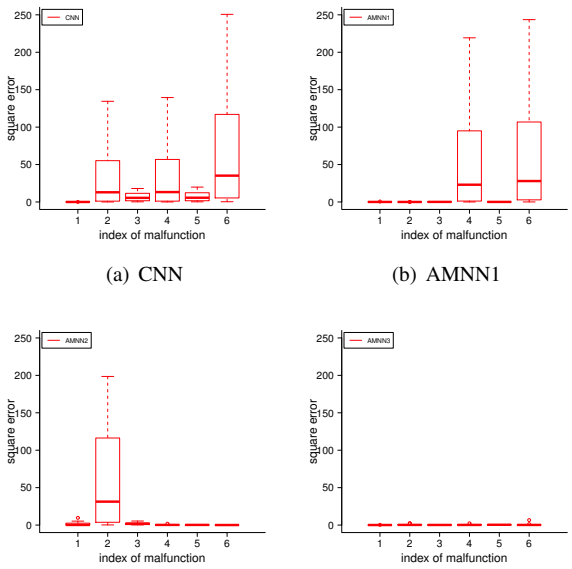


Fig. 13. This plot shows the fitness value of the best individuals in each generation of the evolution for the CNN and AMNNs training schemes.

shown in Figure 14(a). For the AMNN1, when only one DOF malfunctions, the system performs well. Surprisingly, it still performs very well when the third and first DOF malfunction (a_1 and a_3 are prohibited). Note that the system is not trained for this case. For the AMNN2 training, the system can handle more malfunction cases (1, 3, 4, 5 and 6). For the AMNN3, in all of the malfunction cases the trained system can perform well. It seems that the more degrees of freedom are restricted during the training, the higher robust the trained system is, at least when dealing with malfunctions.



(a) CNN (b) AMNN1

(c) AMNN2 (d) AMNN3

Fig. 14. This plot shows the performance of the trained system when malfunctions occur for the CNN and AMNNs training. 1/2/3: only the third/second/first DOF malfunctions. 4: the third and second DOF malfunction; 5: the third and first DOF malfunction; 6: the second and first DOF malfunction.

Generally saying, it is hard to envisage what would be the success of each training before analysing the statistical data. Moreover, it is rather unclear how to relate a success or failure of one specific training with respect to a specific DOF (e.g., the failure of AMNN2 with respect to the second DOF). Clearly, the non-linearity of the equations, the tansig function etc. hinders success in such a forecast.

3) Environmental Change for example 2: Here, the environmental change is simulated by changing the powers or coefficients of the trained function, i.e., the trained system has to approximate different, untrained-for, functions. Figure 15 and Figure 16 show the performance of the CNN and AMNNs training schemes when the powers and coefficients of the target function are changing, respectively. The x coordinates are the same as those in the 10 training points, but the y coordinates have changed depending on the testing functions.

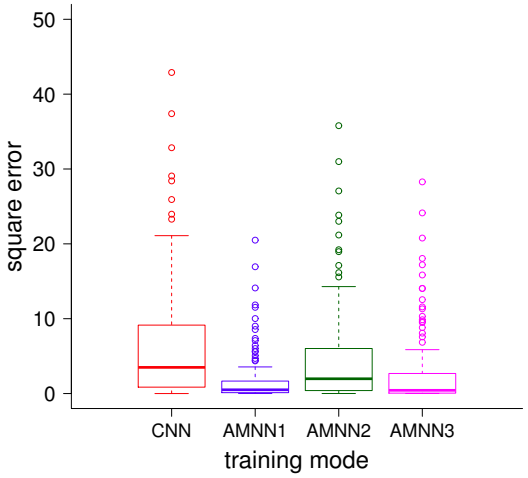


Fig. 15. This plot shows the performance of the CNN and AMNNs training when changing the powers of the target function.

For the case of changing powers, the second order of the target function is changing from $[1.0, 1.2, \cdots, 3.0]$, and the first order is changing from $[0.0, 0.2, \cdots, 2.0]$. We deleted the combination of 2.0 and 1.0, which corresponds to the same target function used for training. For each new testing function, we tested the performance of CNN and AMNNs using the new inputs. To approximate each point in the testing function, we selected the best mode operation that can obtain the least square error. For the CNN, we only have one operation mode. For the AMNN1, AMNN2 and AMNN3 training, the number of selected operation modes is four (one for the non-restricted mode and three for the one DOF restriction modes), four (one for the non-restricted mode and three for the two DOF restriction modes), and seven (one for the non-restricted mode three for the one DOF restriction and three for the two DOF restriction modes), respectively. For each testing function, we assigned an fitness, which is the average square errors for all the 10 points. The case of changing coefficients is similar. In this case, the coefficients corresponding to the second order is changing from $[2.0, 2.2, \cdots, 4.0]$, and the

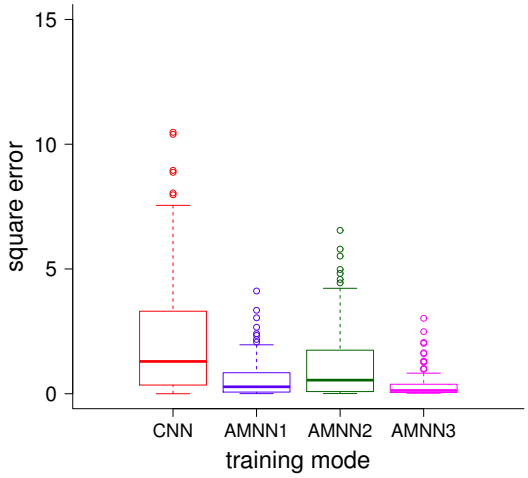


Fig. 16. This plot shows the performance of the CNN and AMNNs training when changing the coefficients of the target function.

coefficients corresponding to the first order is changing from $[1.0, 1.2, \cdots, 3.0]$. We also deleted the combination of 3.0 and 2.0, which corresponds to the original, trained-for, function.

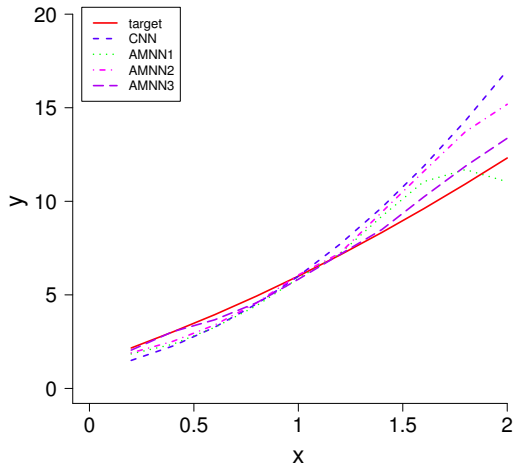


Fig. 17. This plot shows an example of how the CNN and AMNNs approximate a new function when changing the powers of the target function.

It is clear that the AMNN'S settings perform significantly better than that of the CNN, when environmental changes occur (two-sided Mann–Whitney test, 5% significance level). The p values are smaller than 0.02. In the case of changing powers and coefficients, the AMNN1 and AMNN3 outperforms the AMNN2 training. Although we find that restricting two degrees of freedom in the training (AMNN2) benefits more on the malfunctions comparing with restricting only one DOF. In the case of environmental changes, AMNN1 benefits more, as it has more degrees of freedom to approximate the new functions. Comparing the performance of AMNN1 and

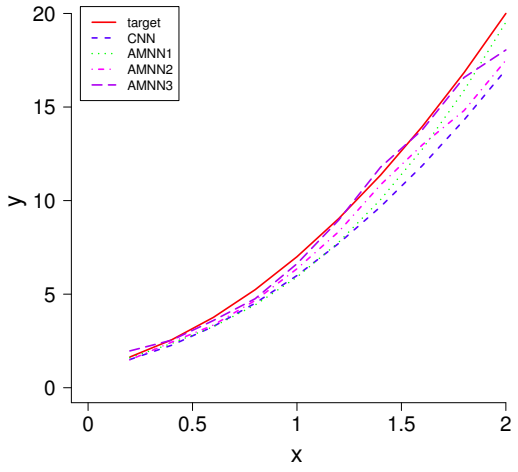


Fig. 18. This plot shows an example of how the CNN and AMNNs approximate a new function when changing the coefficients of the target function.

AMNN3, there is no significant difference in the case of changing the power, but AMNN3 performs significantly better in the case of changing the coefficients. From the results, we can see that the system trained with all combinations of restricted modes possesses the highest robustness.

Figure 17 and Figure 18 show an example of the two cases of environmental changes. The testing functions are still $y = 3 * x^{1.5} + 2 * x^{0.5} + 1$ and $y = 3.5 * x^2 + 2.5 * x^1 + 1$ respectively. As we can see, since the AMNNs have the advantage of choosing different operation modes to approximate the new functions, they perform better than the CNN scheme.

V. Discussion and Future work

The Mechanical Cognitivization idea has been further tested and elaborated here. Although the purpose of MC is the enhancement of robots' robustness, functions are still facilitated here. Clearly, using functions to prove a concept that is to enhance robustness of robots, which are not modeled as functions, requires a leap of faith on behalf of the reader. Nevertheless, much progress on the utilization of neural nets was instigated by using functions and therefore they are used as a fundamental base for our planned study. It has been shown in the paper that learning that includes both non-restricted modes as well as restricted modes, enhances robustness to environmental changes and to malfunctions. Although the training is done on the same training set as trained by the common scheme, the MC based schemes attain improved robustness. The improved robustness is attained through embedding a set of flags that open the way for deliberately restricting modes to attain improved performances. The enhanced robustness comes on the expense of extended training time. In the current paper, an insight is gained on the influence of choosing the restricted modes to be trained with. From analysing the results, it is evident that as more restricted modes are trained for, so does the robustness improves. As for future work, clearly the next step would be proving the MC idea by utilizing a mechanical system such as manipulator/robot. Further research should take place for finding an optimization approach that will optimize the number of DOF (size of n) such that the robustness would be maximized. A multi objective approach could be also taken in order to maximize robustness and training time. Due to the contradiction among these objectives that was highlighted in the current paper, a trade-off set might be found. The utilization of neural nets as the controller's architecture should be also revisited and more sophisticated ones should be considered to serve as the cognitive, learning entity.

VI. Acknowledgements

This research was supported by a Marie Curie International Research Staff Exchange Scheme Fellowship within the 7th European Community Framework Programme.

References

[1] G. Avigad and A. Weiss, "Mechanical Cognitivization," in *Proceedings of the 6th International Conference on Advanced Cognitive Technologies and Applications (COGNITIVE 2014)*. IARIA, 2014, pp. 116–119.

[2] A. Newell, "The knowledge level: presidential address," *AI magazine*, vol. 2, no. 2, p. 1, 1981.

[3] P. S. Rosenbloom, J. E. Laird, and A. E. Newell, *The Soar papers: Research on integrated intelligence, Vols. 1 & 2*. The MIT Press, 1993.

[4] D. E. Kieras and D. E. Meyer, "An overview of the EPIC architecture for cognition and performance with application to human-computer interaction," *Human-computer interaction*, vol. 12, no. 4, pp. 391–438, 1997.

[5] P. Langley, "An adaptive architecture for physical agents," in *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE, 2005, pp. 18–25.

[6] D. Vernon, G. Metta, and G. Sandini, "A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 2, pp. 151–180, 2007.

[7] L. A. Shapiro, "The mind incarnate," 2004.

[8] ——, *Embodied cognition*. Routledge London, 2011.

[9] M. Shanahan, "Consciousness, Emotion, and Imagination," in *Proceedings of the 2005 AISB Workshop: Next Generation Approaches to Machine Consciousness*, 2005, pp. 26–35.

[10] C. Burghart, R. Mikut, R. Stiefelhagen, T. Asfour, H. Holzapfel, P. Steinhaus, and R. Dillmann, "A cognitive architecture for a humanoid robot: A first approach," in *Proceedings of the 5th IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2005, pp. 357–362.

[11] P. Haazebroek, S. Van Dantzig, and B. Hommel, "A computational model of perception and action for cognitive robotics," *Cognitive processing*, vol. 12, no. 4, pp. 355–365, 2011.

[12] J. R. Simon and A. P. Rudell, "Auditory SR compatibility: the effect of an irrelevant cue on information processing." *Journal of Applied Psychology*, vol. 51, no. 3, p. 300, 1967.

[13] R. Pfeifer, F. Iida, and G. Gómez, "Morphological computation for adaptive behavior and cognition," in *International Congress Series*, vol. 1291. Elsevier, 2006, pp. 22–29.

[14] R. Pfeifer and G. Gómez, "Morphological computationconnecting brain, body, and environment," in *Creating Brain-Like Intelligence*. Springer, 2009, pp. 66–83.

[15] T. M. Kubow and R. J. Full, "The role of the mechanical system in control: a hypothesis of selfstabilization in hexapedal runners," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 354, no. 1385, pp. 849–861, 1999.

[16] H. Yokoi, A. H. Arieta, R. Katoh, W. Yu, I. Watanabe, and M. Maruishi, "Mutual adaptation in a prosthetics application," in *Embodied Artificial Intelligence*. Springer, 2004, pp. 146–159.

[17] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida, "Cognitive developmental robotics: a survey," *IEEE Transactions on Autonomous Mental Development*, vol. 1, no. 1, pp. 12–34, 2009.

[18] M. H. Lee, Q. Meng, and F. Chao, "Developmental learning for autonomous robots," *Robotics and Autonomous Systems*, vol. 55, no. 9, pp. 750–759, 2007.
[19] M. Hülse, S. McBride, and M. Lee, "Fast learning mapping schemes for robotic handeye coordination," *Cognitive Computation*, vol. 2, no. 1, pp. 1–16, 2010.
[20] J. Law, M. Lee, M. Hülse, and A. Tomassetti, "The infant development timeline and its application to robot shaping," *Adaptive Behavior*, vol. 19, no. 5, pp. 335–358, 2011.
[21] H.-G. Beyer, *The theory of evolution strategies*. Springer, 2001.

Representing and Publishing Cyber Forensic Data and its Provenance Metadata: From Open to Closed Consumption

Tamer Fares Gayed, Hakim Lounis
Dépt. d'Informatique
Université du Québec à Montréal
Succursale Centre-ville, H3C 3P8,
Montréal, Canada
gayed.tamer@courrier.uqam.ca lounis.hakim@uqam.ca

Moncef Bari
Dépt. de Didactique
Université du Québec à Montréal
Succursale Centre-ville, H3C 3P8,
Montréal, Canada
bari.moncef@uqam.ca

***Abstract*—Role players of any forensic investigation process record chronologically all forensic data resulted from their investigation, in order to be presented to the juries in the court of law. When such results are recorded and posted, they are called chain of custodies (*CoCs*). The forensic data provided within these documents play a vital role in the process of forensic investigation, because they answer questions about how evidences are collected, transported, analyzed, and preserved since their seizure through their production in court. Provenance metadata accompany these forensic data to answer questions about the origin of these data and build trustworthy between role players and juries in order to make the tangible *CoCs* admissible in the court of law. Nowadays, with the advent of the digital age, the forensic investigation is not only applied to physical crime, but also on digital evidences. The forensic data and their metadata presented in these tangible documents need also to undergo a radical transformation from paper to electronic data in order to accommodate this evolution. *CoCs* should be also readable and consumable not only by human but also by machines. The semantic web is a fertile land to represent and manage the tangible *CoCs*, because it uses web principles known as Linked Data Principles (LDP), which provide useful information in Resource Description Framework (RDF) format upon Unified Resource Identifiers (URI) resolution. In addition, it includes different provenance vocabularies that can be useful to express the forensic metadata. Generally, the power of LDP resides in publishing data publicly without any access restriction on the web. However, the openness of forensic data and their metadata should not be the same case. They should obey some access restriction in order to be shared only between role players and juries. Public Key Infrastructure (PKI) can be applied to restrict the access to some or all resources of represented data and bends the LDP from open to closed consumption, while maintaining the resolution of such restricted resources. Juries in turn will consume the restricted represented data using different LDP consumption applications. This paper provides the complete framework explaining how forensic and provenance data are represented and published using LDP, and how PKI can be used to restrict these data/resources in order to be shared in a closed scale. Evaluation of the framework using several empirical experimentations will not be on the scope of this paper.**

Keywords-Linked Open Data, Linked Data Principles, Linked Closed Data, Public Key Infrastructure, Digital Certificates, Cyber Forensics, Chain of Custody.

I. INTRODUCTION

The history of forensic investigation task dates back thousands of years. This task is concentrating to gather and examine evidences about the past, in order to prosecute in the future the criminal in the court of law. With the advent of Information and Communication Technology (ICT), forensic investigation is not only concentrated on physical crime, but also on the digital evidences. This emerged a new type of forensic investigation known by computer/cyber/digital forensic. It combines computer science concepts including computer architecture, operating systems, file systems, software engineering, and computer networking, as well as legal procedures. At the most basic level, the digital forensic process has three major phases: extraction, analysis, and presentation. Extraction phase (i.e., it is also known as acquisition) saves the state of the digital source (e.g., laptop, desktop, computers, mobile phones, or any other digital devices) and creates an image by saving all digital values so it can be later analyzed [1]. Analysis phase takes the acquired data (e.g., file and directory contents and recovering deleted contents) and examines it to identify pieces of evidence, and draws conclusions based on the evidences that were found. During presentation phase, the audience is typically the judges; in this phase, the conclusion and corresponding evidence from the investigation analysis are presented to them [2][3].

However, there exist others models of cyber forensic process, each of them relies upon reaching a consensus about how to describe digital forensics and evidences [4][5]. Investigation models are numerous. Many works were provided to explain and compare such models [6][7][8][9]. Table I shows the current digital forensic models. Each row of the table presents the name of the digital forensic process model, while the columns present the processes included in each of these models [5][10].

The role players such as first responders, investigators, expert witnesses, prosecutors, police officer, etc. may be assigned one or more phase in the forensic process. They are those who are responsible to create and record their own investigation results and post them in tangible documents.

TABLE I. DIGITAL FORENSIC PROCESS MODELS [5]

	Acquire	Authenticate	Analyze	Collection	Examination	Reporting	Recognition	Identification	Individualisation	Reconstruction	Preservation	Classification	Presentation	Decision	Preparation	Approach Strategy	Returning Evidence	Awareness	Authorization	Planning	Notification	Transportation	Storage	Hypothosis	Proof/defence	Dissemination
Kruse	*	*	*																							
USDOJ			*	*	*	*																				
Casey							*			*	*	*														
DFRWS			*	*	*			*			*		*	*												
Reith			*	*	*			*			*		*		*	*	*									
Ciardhuain				*	*								*					*	*	*	*	*	*	*	*	*

These documents are known by chain of custodies, as they record all collected evidences (forensic data) in their chronological order, in order to avoid later allegations of tampering with such evidences. *CoC* considered as a testimony document and one of the most essential parts of any forensic investigation process [6], because it provides useful information about the evidences studied through different forensic process by answering 5Ws and 1H questions. The 5 Ws are the When, Who, Where, Why, What and the 1 H is the How. *CoC* must include documentation containing answers to these questions. For example:

- Who came into contact, handled, and discovered the digital evidence?
- What procedures were performed on the evidence?
- When the digital evidence is discovered, accessed, examined, or transferred?
- Where was digital evidence discovered, collected, handled, stored, and examined?
- Why the evidence was collected?
- How was the digital evidence collected, used, and stored?

Once such questions are answered for each phase in the forensic process, players will have a reliable *CoC*, which can be then admitted by the judges in the court of law.

Reliability on information is not enough to admit the *CoC* in the court of law. Trustworthiness is also required. On the level of data, it occurs when receivers (i.e., juries) ensure from the *origin* of data that the senders (i.e., role players) sent to him, and this will be realized through different provenance vocabularies. On the level of players, trustworthiness occurs when receivers ensure from the *identity* of the senders (i.e., called also repudiation), and this will be realized through the PKI. Provenance of information related to data and identities of players are crucial to guarantee the trustworthiness and confidence that the role players provided to the juries.

Thus, each *CoC* document contains not only forensic data, but also data describing the origin of this data (i.e., the forensic data presented in the *CoC* will be accompanied by metadata). Metadata is the data that describes other data. Thus, forensic information is responsible to answer the 5Ws and 1H questions related to the forensic investigation, while provenance information is responsible to answer questions about the origin of these forensic data. For example:

- Who published/created the data?
- What is the published date?
- Where this data is initially published/created?
- When/Why the data is published?
- How the data is published?

The questions of forensic data may differ from one phase to another in the forensic process. Their questions must be posed separately over each phase of the forensics process (i.e., 'What' question, of the collection phase is not the same as the 'What' for the identification phase). For example, the Kruse model (see Table I, first row) has 3 forensics phases, thus, it should have 3 different *CoCs* [11]. Nevertheless, most of works provided in the forensics process globalize the 5Ws and 1H questions once over the whole forensics process [7][9].

Nowadays, with the advent of digital age, *CoCs* should be transformed from tangible document to electronic form consumable not only by the human but also by the machine. There are three main motivations do this task [12]:

- *Motivation 1*: cyber forensics is a daily growing field that requires the accommodation on the continuous changes of digital technologies as well as its tangible documents (i.e., concurrency with the knowledge management). Thus, tangible *CoCs* and all their contents (i.e., victim information and forensics information) must also undergo a radical transformation from paper to machine-readable format in order to accommodate this continuous evolution.
- *Motivation 2*: judges' awareness and understanding the digital evidences are not enough to evaluate and take the proper decision about the digital evidence. Juries need to know more concerning the evidences in hand. One of the proposed solutions is to organize a syllabus and training program to educate the juries the field of ICT [13]. The authors argue against this solution direction, because it will not be an easy task to teach juries with their juridical positions, the different concepts of ICT. The authors propose a solution offering the ability to the juries to navigate, discover (dereference) and execute different queries on the represented information.
- *Motivation 3: CoC*s play vital role in the investigation process and due to this fact, it must be maintained and managed throughout the investigation process, in order to preserve its integrity, especially when the evidence has digital nature. However, if the *CoC* is not well maintained and the suspect was guilty, a lawyer/defense can argue that the *CoC* was not properly established and casting doubt on the damning of the acquired evidence. A security mechanism should be integrated with the represented data to keep its

integrity to limit and control its access to only the authorized people.

Semantic web will be a flexible solution for this task because it provides several semantic markup languages such as Resource Description Framework (RDF) [14], RDF Scheme (RDFS) [15], and Web Ontology Language (OWL) [16] that are used to represent different data and knowledge. In addition, the semantic web is rich with different provenance vocabularies [17], such as Dublin Core (DC) [18], Friend of a Friend (FOAF) [19], and Proof Markup Language (PML) [20] that can be used to (im)prove the *CoC* by answering the 5Ws and the 1H questions [3].

Furthermore, the semantic web today is the web of data, which is not just concentrated on the interrelation between web documents, but also between the raw data within these documents. This interrelation of data is based on three aspects known as the LDP or the technology stack. The latter contains Unified Resource Locators/Identifiers (URL/URI) [21], Hypertext Transfer Protocol (HTTP) [22], and RDF [14]. Simply, this stack is used to publish data in a structured way that facilitates their consumption through representing and naming different resources using URL/Unified Resource Identifier (URI).

The Linking Open Data (LOD) project is the most visible project using this technology stack (URLs, HTTP, and RDF) that converts existing open license and provenance data on the web into RDF according to the LDP [23][24]. Thus, the LOD publish open data on the web without access restriction. However, this will not be feasible in a context where only role players and juries need to publish and consume, respectively, the represented data in a small scale.

Represented URI/URL resources of *CoCs (e-CoCs)* need to obey then some access restriction, where a specific set of people are those who are authorized to access such resources. LDP should be bended to realize the adaptation of publishing and consuming the resources on a small scale without loosing the resolvability feature of these resources. Thus, a compromise question arises in this case, how we can realize the access restriction over certain URI/URL resources while keeping the resolvability feature of the same resources. In addition, this question brings out a new era of research called the Linked Closed Data (LCD) [25], where the publisher would take step of imposing access restrictions to protect his information [25][26][27]. Finally, the represented resources will be closed and shared only between role players and juries. The latter can consume resources using different pattern consumption; whether by browsing, crawling, querying, or reasoning on the represented data.

This paper extends the work published in [1]. In this work, a framework solution called Cyber Forensic-*CoC* (CF-*CoC*) has been provided (see Figure 1). One of the layers was the PKI layer that was used to bend the LDP from LOD to LCD. The current work resumes the work published in [1], by depicting all the layers together and by clarifying how the PKI are applied to restrict the publication and consumption of forensic data and provenance metadata.

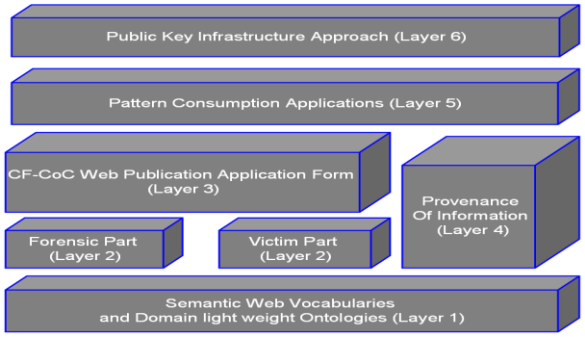


Figure 1. Cyber Forensics-Chain of Custody (CF-*CoC*) Framework

It also explains the two remaining layers of the CF-*CoC* framework (i.e., provenance metadata layer and consumption layer) that were not provided and published in our recent works [1][3][10][11][12][28].

Furthermore, as mentioned before, this work can be used to argue against the solution proposed in [13] concerning the judges' awareness and understanding of the digital evidence. This solution helps the juries to understand the field of ICT. The aim of this research is the construction of a system, offering the ability for the role player to record and publish electronically their forensic investigation and for the juries to navigate, discover (i.e., dereference), and execute different queries on the represented information in order to understand the case in hand.

This paper is organized as follows: Section II is the state of the art that depicts different disciplines related to the CF-*CoC* framework. Section III states the advantages of using LDP to represent *CoC*, Section IV provides the research problem, Section V depicts the CF-*CoC* framework and system. Finally, Section VI provides the conclusion and future works.

II. STATE OF THE ART

The state of the art related to this framework goes over different disciplines such as semantic web, cyber forensics, provenance of information, and security, so the state of the art in this section will have different facets. Each facet discusses the related works of each discipline apart (see Figure 2).

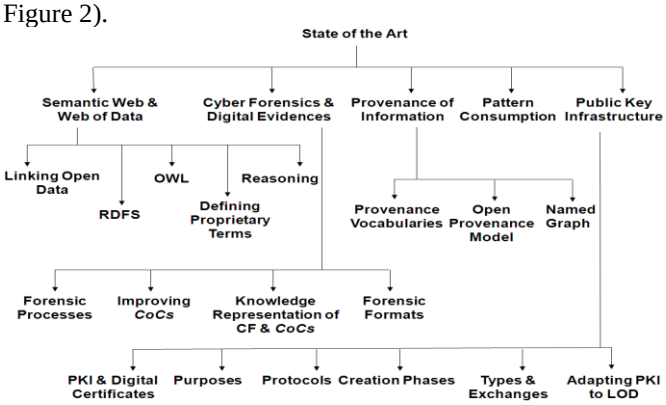


Figure 2. State of the Art related to CF-*CoC* Framework

A. Semantic Web and Web of Data

Semantic web is an extension of the current web (i.e., from document to data) [27][29], designed to represent information in a machine readable format by introducing RDF model [14] to describe the meaning of data and allows them to be shared on the web in a flexible way. The classical way for publishing documents on the web is just naming these documents using URI and hypertext links. This fact allows the consumer to navigate over the information on the web using a web browser application and crawling the information by typing keywords in a search engine that is working using the support of HTTP protocol. This is called the web of documents.

With the same analogy, entities and contents (i.e., data) within documents can be linked between each others using typed linked and with the same principles used by the web (i.e., web aspects). This is called the web of data.

Nowadays, the main aim of the semantic web is to publish data on the web in a standard structure, and manageable format [35]. Tim Berners-Lee outlined the principles of publishing data on the web. These principles known as Linked Data Principles (i.e., LD principles) [24] [27]:

- Use URI as names for things.
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information using the standards (RDF, SPARQL).
- Include RDF statements that link to other URIs so that they can discover related things.

According to the W3C recommendation [14], RDF is a foundation for encoding, exchange, and reuse of structured metadata. It can be serialized using different languages (e.g., RDF/XML [30], Turtle [31], RDFa [32], N-Triples [33], N3 [34]). RDF consists of three slots called triples: resource, property, and object. In addition, resources are entities retrieved from the web (e.g., persons, places, web documents, pictures, abstract concepts, etc.). RDF resources are represented by uniform resource identifiers (URIs) of which URLs are a subset.

After the resources are identified using URIs, they will be connected using RDF links, creating a global data graph that spans data sources and enable the resolvability of such resources to a new data source. The LOD cloud project has been constructed upon this basic structure.

1) Linked Open Data

The Linked Open Data (LOD) project is the most visible project using this technology stack (URLs, HTTP, and RDF) and converts existing open license data on the web into RDF according to the LDP [35][24] (see Figure 3).

The LOD is based on the LDP, where URI resources are linked using typed RDF links to other resources within the same or to other data set. Two types of links can be used; links to navigate forward and others to navigate backward between resources. For example, if we have an RDF triple connecting two resources x and y, and we need to move

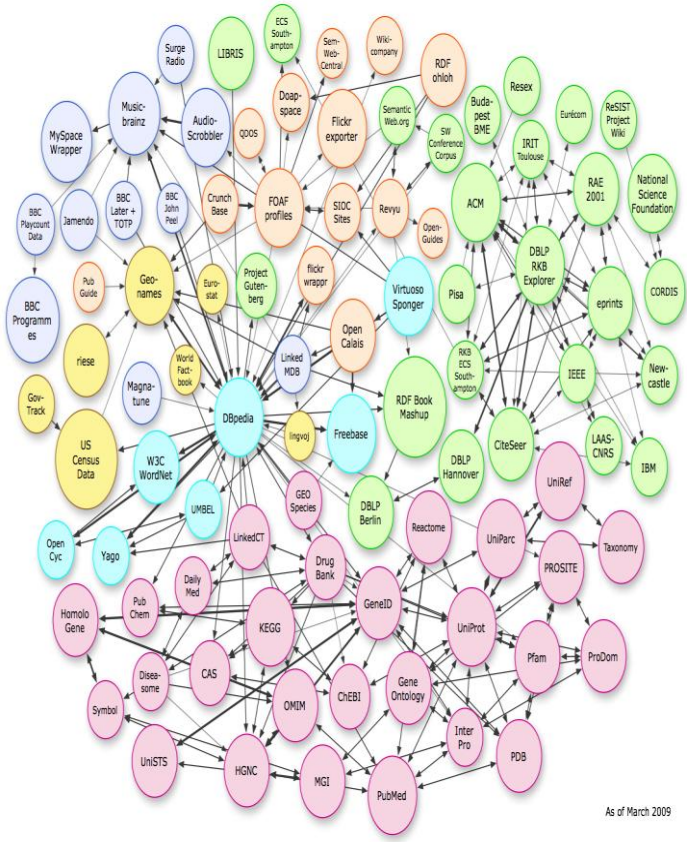


Figure 3. Linking open data cloud diagram

forward from x to y, then this RDF triple should appear in the document describing the resource y. This triple is then called incoming link, because it allows to navigate back to resource x. Same case, for the outcoming link, where the RDF triple should appear in the document describing the resource x and allows to navigate forward to resource y [36]. Figure 3 shows the LOD cloud diagram, where each links exists between items in the two connected data sets. Some data sets are connected together using whether, the outcoming links, the incoming links, or both.

The LOD project created a shift in the semantic web community. Instead, the concern was on the ontologies for their own sake and semantic, it becomes on the web aspects (how to publish and consume data on the web).

Ontologies are used then to foster and serve the semantic interoperability between parts that want to exchange such data. These are known as lightweight ontologies [37] that use the full advantages of semantic web technologies, minimum OWL constructs, and reuse existing RDF vocabularies wherever possible.

Resources have properties (attributes) that admit a certain range of values or that are attached to another resource. The object can be a literal value or a resource.

While RDF provides the model and syntax for describing resources, it does not define the meaning of those resources. That is where other technologies such as RDF Schema

(RDFS) come in [36]. RDFS specifies extensions to RDF that are used to define the common vocabularies in RDF metadata statement and enables specification of schema knowledge. It develops classes for both resources and properties. However, RDFS is limited to a subclass hierarchy and a property hierarchy with domain and range definitions of these properties. RDFS limitations are range restrictions, disability of expressing disjoitness between classes, combination between classes, cardinality restriction, and characteristics of properties [38].

Thus, RDF is the standard format to create LD and it is sufficient to use the constructors of RDFS and a little feature of OWL to represent data in LD structure. Combination of constructors from both vocabularies (i.e., RDFS and OWL) represents the lightweight ontology of RDF and LD. This is known by RDFS++. Next subsections highlight all the RDFS constructors and some OWL primitive constructors that will be used to construct the first two layers of CF-*CoC* framework.

The RDFS and OWL constructors are classified according the term type (i.e., *rdfs:class*, or a property *owl:objectProperty*). This definition takes place before the term will be used (i.e., before its publication, T-Box). Later, the defined terms are used to describe and publish different data (A-Box, Assertion Box) [38]. The type of the term also determines its slot position during publication.

2) RDFS Constructors

The RDFS constructors are used to define terms and their relationships. Consider the term in question is named *X* (see Table II).

TABLE II. RDFS CONSTRUCTORS FOR PROPERTY AND CLASS TERMS

If *X* is a term of type (*rdf : type*) Property (*rdfs : Property* / *owl : ObjectProperty*)	
rdfs : subPropertyOf	When the term *X* is of type *property* it can be also a sub property of another *property* term. The *subPropertyOf* of a property term is a term of type *Property*
rdfs : range	The *range* of a property term is always a *Class*. A *range* of a property term *X* states that the **object** slot of the *X* (i.e., where *X* is a predicate, because *X* is a *property)*, interpreted by a reasoners as an instance of said *range* of *X*
rdfs : domain	The *domain* of a property term is always a *Class*. A *domain* of a property term *X* states that the **subject** slot of the *X* (i.e., where *X* is a predicate, because *X* is a *property)*, interpreted by a reasoners as an instance of said *domain* of *X*
If *X* is a term of type (*rdf : type*) Class (*rdfs : Class*)	
rdfs : subClassOf	When the term *X* is of type *Class*, it can bee also a sub class of another *Class* term. The *subClassOf* of a property term is a term of type *Class*
Common Constructors between *Property* and *Class* terms	
rdfs : comment	Any term should have a *comment*. A *comment* is used to provide a human-readable description of a resource. *Comment* is an instance of *rdf : Property*
rdfs : label	Any term should have a *label*. A *label* is used to provide a human-readable name for a resource. *Label* is an instance of *rdf : Property*

3) OWL Constructors

The primitive selected from the OWL are mainly used to map between class and property terms (see Table III).

TABLE III. OWL CONSTRUCTORS FOR PROPERTY AND CLASS TERMS

If *X* is a term of type (*rdf : type*) Property (*rdfs : Property* / *owl : ObjectProperty*)	
owl : equivalentProperty	This constructor is used to map between two terms of type *Property*
owl : InverseProperty	This constructor is used to state that one property is the inverse of another. It is use to describe inverse relation between properties (i.e., exactly like the passive voice in the grammar)
owl : InverseFunctionalProperty	When the type (*rdf:type*) of a *property* term *X* is defined to be of *InverseFuntionalProperty*, Whenever *X property* is used as a predicate in a triple, its **object** will have **one** and **only one subject**. Thus, each object should be able to uniquely identify a subject. This constructor is a sub class of *owl : objectProperty*
owl : FunctionalProperty	Same idea as the last constructor, but here, when *X* is defined to be of type *FunctionalProperty*, each **subject,** where *X* is a predicate, can have at most one **object.** This constructor is a subclass of *rdf : property*
If *X* is a term of type (*rdf : type*) Class (*rdfs : Class*)	
owl : equivalentClass	This constructor is used to map between two terms of type *Class*
Common Constructors between *Property* and *Class* terms	
owl : sameas	Two URI terms can be mapped together using the *sameas* constructor. This constructor indicates that these two terms actually refer to the same thing. It can be used as well to map between two ontologies.

These constructors in Tables II and III, are used to publish data on web. Publication of terms on the web passes by three steps. It starts with identifying terms in the domain of interest. These terms are the things whose properties and relationships will be used later in the publication of data.

The identification (selection) is achieved through the descriptions of different processes and tasks performed within each forensic phase [39][40]. The identified terms are also called custom or proprietary terms.

Second step, the identified terms are defined using different constructors of RDFS [15] and OWL [16], and uniquely named by HTTP URIs. The defined terms are then used to publish different information.

4) Defining Proprietary terms

The existing terms defined by the vocabularies of the semantic web are not enough to describe all domains. Sometimes, there are no existing ontologies (vocabularies) containing terms describing a particular data set (e.g., cyber forensics). Some domains are new and others are still in their infancy. This is the case of CF, where it is scarce to find forensic terms or well-known vocabularies describing it, because this domain is still in its infancy and development. Thus, new proprietary terms need to be defined and developed in a dedicated vocabulary, applying the features of RDFS [15] and OWL [16] to describe this particular data set. However, before creating a new custom term, some aspects (criterions) should be taken into consideration [41]:

- Search for terms from widely used vocabularies that could be reused to describe the domain in interest. If the widely deployed vocabularies do not provide the required terms to describe such domain, so new terms should be defined as proprietary terms.
- When you define a new term, you need to have a namespace that you own and control (i.e., unique namespace), in order to mint your new terms to this domain/namespace.
- When you create new terms, you have to map these terms to those in existing vocabularies.
- Apply the LDP to your new terms by using the web technology stack (HTTP, URL, and RDF) and this

task takes place along the publication process, starting from the identification of terms until their publication.
- Label and comment each term you create.
- If your term is a property (predicate), you have to define its *domain* and *range* using the constructors of RDFS and do not overload your new term with ontological axioms.
- If at later time you discover that another term was enough, an RDF link should be set between the new created term and the existing one.

Although, there exist different guides to publish terms, the process of selecting and identifying them is remaining a subjective task and depends on the term creator (i.e., we may have two creators selecting and identifying two different terms describing the same concept in the real world). This does not affect the quality of terms being published, because the LDP on the web of data make them self-descriptiveness. The latter advantage is due to two reasons:

- LDP with naming using HTTP/URIs, offer a dereferenceable nature to the term, so that any LD consumption applications can look up the RDFS/OWL definitions and retrieve more information about such term [42].
- LDP with some schema constructors (i.e., OWL) can map a new term to existing terms from well-defined vocabularies in the form of RDF links [43].

The most related work to define an ontology in *CF*, was published in [44], where an ontological model (i.e., with small 'o') was created for outlining CF tracks in the education process. Its aim was only to construct a hierarchical structure for classification of certification domains (i.e., the best convenient vocabulary to be used by the web of data to construct such type of ontology is the Simple Knowledge Organization System – SKOS [45]). Thus, CF is a domain that requires the definition of new proprietary terms. The CF-*CoC* framework provided in this paper aids the role player to represent *CoC* by defining new proprietary terms and publish such information on the web of data in RDF format.

Any forensic process contains a set of phases, where each phase is assigned to one or more role player. Thus, the number of forensic phases and how many role players assigned to each phase determine the total number of role players participated in the forensic investigation. Each forensic phase contains a set of forensic tasks; each task can be described using a set of terms representing the forensic information.

Before discussing how the role player can use the CF-*CoC* to generate the *e-CoC* (see Section V), it is necessary to explain how the forensic information can be ontologically mapped. As shown in Figure 4, each forensic phase will have a corresponding lightweight ontology. Each lightweight ontology has a set of *n* categories, which will be equivalent to *n* forensic tasks. A category in the vocabulary should be described using a set of *m* terms. These terms are the proprietary terms describing a forensic task.

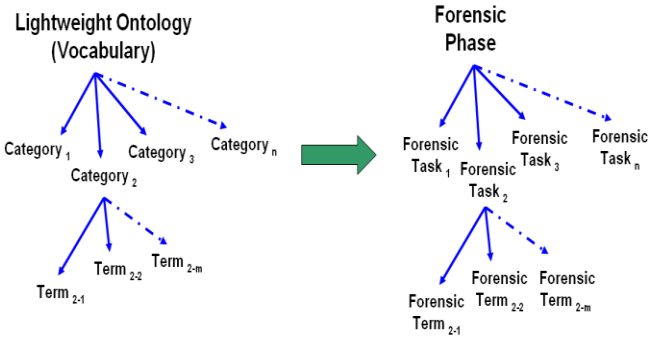


Figure 4. Mapping between Ontological Concepts and a Cyber Forensic Phase [28]

This work considered the preservation task of the acquisition phase imported from Kruse model [46], as an example to elaborate the idea of creating lightweight ontology with new proprietary forensic terms.

5) Reasoning

As mentioned in Section II.A.1, lightweight ontology of LD is a combination of RDFS constructors and some primitive of OWL. Inference is a derivation of logical conclusion from premises known or assumed to be true. Reasoning is a process to extract new information from existing information stored in a knowledge base. For the LD, the knowledge base is the RDF triples store.

RDFS and OWL contains set of inference rules related to their constructors. This section discusses the rules of RDFS constructors, and some rules of OWL (i.e., those that are primitives and used to describe the LD). Table IV depicts the rules of the most used constructors of both vocabularies (i.e., RDFS, and OWL).

TABLE IV. RULES AND ENTAILMENTS OF RDFS AND OWL [16][47]

Constructor Name	Rules and Entailments
rdfs:subClassOf	***subClassOf* is transitive when:** *(A, rdfs:subClassOf, B), (B, rdfs:subclassOf, C) => (A, rdfs:subClassOf, C)* **Another Entailment rule of *subClassOf:*** *(a, rdf:type, A),(A,rdfs:subClassOf B) => (a, rdf:type B)*
rdfs:subPropertyOf	***subPropertyOf* is transitive when:** *(a, p, b), (p, rdfs:subPropertyOf q) => (a, q, b)* **Another Entailment rule of *subPropertyOf:*** *(a, p , b), (p rdf:type rdf:Property)*
rdfs:domain	*(p, rdfs:domain, A), (a, p, x) => (a, rdf:type, A)*
rdfs:range	*(p, rdfs:range, A), (x, p, a) => (a, rdf:type A)*
owl: FunctionalProperty	*If a property p is tagged as FunctionalProperty then all x,y, and z : p(x,y) and p(x,z) => y=z*
owl: InverseFunctionalProperty	*If a property p is tagged as InverseFunctionalProperty then all x,y and z: p(y,x) and p(z,x) => y=z*
owl: inverseof	*If a property p1, is tagged as the owl:inverseof p2, then for all x and y: p1(x,y) iff p2(y,x)*

B. Cyber Forensics Processes and Digital Evidences

Second discipline in the state of the art section is related to the cyber forensic and digital evidences. Despite the infancy of the CF field, many works have been provided related to the forensic processes, *CoC*, and forensic formats.

1) First Category : Forensic Processes

The works provided under this category concentrated on the creation of different forensics processes. Different Digital Forensics Process Models (DFPM) has been proposed since 2000 (e.g., Kruse [46], the United State Department of Justice (USDOJ) [6], Casey [7], Digital Forensics Research Workshop (DFRW) [48], and Ciarhuin [8]) to assist the players of investigations process reaching conclusions upon completion of the investigation.

As been mentioned in Section I, investigation models are numerous. Many works were provided to explain and compare such models [5][6][7][8][9] (see Table I). Some phases from different forensics models may have unique technical requirement but they differ only on their names [49]. The work presented by Yussof et al. [9] underlines 46 phases from 15 selected investigation models that have been produced throughout 1995 to 2010, and then identifies the commonly shared processes between these models.

Kruse model is a model that encompasses three major phases of any forensic investigation. The three phases are acquisition of the evidence, authentication of the recovered evidence, and analysis of the evidence. The next three paragraphs explain briefly each phase apart:

- ***Acquisition:*** this phase is about acquiring digital evidences from digital suspected devices (e.g., small-scale devices, large-scale devices, etc.). It contains three forensics tasks: state preservation, recovering, and copying. The role player of this phase is the first responder [6][9].
    - *State preservation*: the first task is saving the state of the digital device under question, by seizing the machine containing the suspected device.
    - *Recovery***:** after seizing the suspected device, the role player tries to recover all deleted files on the device, especially the system files that records valuable details about this suspected device.
    - *Copy:* after recovering the deleted files, the first responder takes copy from the suspected device to avoid tampering and alteration.
- ***Authentication:*** it is the process of ensuring that the acquired evidence has not been altered and kept its integrity since the time it was extracted, to the time it was transmitted, and stored by an authorized source [39]. Any change to the evidence will render the evidence inadmissible in the court. Investigators authenticate the digital media by generating a checksum (Hash) of it contents (i.e., using the MD5, SHA, and CRC algorithms). Checksum is like an electronic fingerprint in that it is almost impossible for two digital media with different data to have the same checksums. The main aim behind this task is showing that the checksums of the seized media (suspected) and the trusted (image) are identical.
- ***Analysis:*** This is the last and most time-consuming step in this model. In this phase, the investigator tries to uncover the wrongdoing of the crime by examining the acquired data such as files and directories in order to identify pieces of evidence and determine their significance and probative value and drawing conclusion based on the evidence found. In [50], the author defined the 3 major categories of evidence that should be considered in the analysis phase:
    - Inculpatory evidence: evidence that supports a given theory
    - Exculpatory evidence: evidence that contradicts a given theory
    - Evidence of tampering: evidence that is used to tamper the system to avoid the correct identification

2) Second Category : Improving the CoCs

Several works are provided in the literature to improve the *CoC*. The work presented in [51] provides the idea of exploiting RDF structure to improve an expansible open format of AFF4. In [52], a conceptual Digital Evidence Management Framework (DEMF) was proposed to implement secure and reliable digital evidence *CoC*. This framework answered the 'who', 'what', 'why', 'when', 'where', and 'how' questions. The 'what' is answered using a fingerprint of evidences. The 'how' is answered using the hash similarity to changes control. The 'who' is answered using the biometric identification and authentication for digital signing. The 'when' is answered using the automatic and trusted time stamping. Finally, the 'where' is answered using the GPS and RFID for geo-location.

Another work in [53] discusses the integrity of *CoC* through the adaptation of hashing algorithm for signing digital evidence put into consideration identity, date, and time of access of digital evidence. The authors provided a valid time stamping provided by a secure third party to sign digital evidence in all stages of the investigation process.

Other published work to (im)prove the *CoC* is based on a hardware solution. SYPRUS Company provides the Hydra PC solution. It is a PC device that provides an entire securely protected, self-contained, and portable device (i.e., connected to the USB Port) that provides high-assurance cryptographic products to protect the confidentiality, integrity, and non-repudiation of digital evidence with highest-strength cryptographic technology [54]. This solution is considered as an indirect (im)proving of the *CoC*

as it preserves the digital evidences from modification and violation.

3) Third Category : Knowledge representation of CF Processes and CoCs

The works of the knowledge representation created in CF concentrate on the representation of the cyber forensics models or on the digital evidences (as indirect improve for the *CoCs*).

An attempt was performed to represent the knowledge discovered during the identification and analysis phase of the investigation process [55]. This attempt uses the Universal Modeling Language (UML) for representing knowledge. It is extended to a unified modeling methodology framework (UMMF) to describe and think about planning, performing, and documenting forensics tasks.

Another work provided in [5] explains how different cyber forensic processes are modeled using the UML. In this work, the behavioral Use Cases and Activity diagrams are presented in order to clarify the limitations of such processes.

A research is also provided in [56] that hypothesis that the formal representational approach will be benefit for the cyber forensics. This work summarized at a fundamental level the nature of digital evidence and digital investigation.

Other works are also presented in [57][58]. They try to improve indirectly the *CoC* through the representation of digital evidences. Both works concentrated mainly on the representation and correlation of the digital evidences and as an indirect consequence, the (im)proving of the *CoC*.

Recently, a new work is provided in [59] to model the forensic process. This work proposed an abstract model for the digital forensic based on the flow-based specification methodology. This methodology is generally used to represent several items such as data, information, or signals using the Flowthing Model (FM) that contains six stages (i.e., arrive, accepted, processed, released, created, and transferred) allowing anyone to draw the system using flow systems.

4) Fourth Category : Forensic Format

Over the last few years, different forensic formats were provided. In 2006, Digital Forensics Research Workshop (DRWS) formed a working group called Common Digital Evidence Storage Format (CDEF) working group for storing digital evidence and associated metadata [60]. CDEF surveyed the following disk image main formats: Advanced Forensics Format (AFF), Encase Expert Witness Format (EWF), Digital Evidence Bag (DEB), gfzip, ProDiscover, and SMART.

Most of these formats can store limited number of metadata, like case name, evidence name, examiner name, date, place, and hash code to assure data integrity [60]. The most commonly used formats are described here. AFF is defined by Garfinkel et al. in [61] as a disk image container, which supports storing arbitrary metadata in single archive, like sector size or device serial number. The EWF format is produced by EnCase's imaging tools. It contains checksums, a hash for verifying the integrity of the contained image, and error information describing bad sectors on the source media.

Later, Tuner's digital evidence bags (DEB) proposed a container for digital crime scene artifacts, metadata, information integrity, and access and usage audit records [62]. However, such format is limited to name/value pairs and makes no provision for attaching semantics to the name. It attempts to replicate key features of physical evidence bags, which are used for traditional evidence capture.

In 2009, Cohen et al. in [63] have observed problems to be corrected in the first version of AFF. They released the AFF4 user specific metadata functionalities. They described the use of distributed evidence management systems AFF4 based on an imaginary company that have offices in two different countries. AFF4 extends the AFF to support multiple data sources, logical evidence, and several others (im)proves such the support of forensic workflow and the storing of arbitrary metadata. Such work explained that the Resource Description Framework (RDF) [14] resources can be exploited with AFF4 in order to (im)prove the forensics process model.

Despite the multiplicity of these forensic formats, role player can use any one of them. Each forensic tool can generate one or more forensic format(s) that can describe specific forensic results (e.g., AFF4 can be generated by EnCase imaging tool, and provide information about the size of digital media, its chunk size, its chunks in segment, etc.). Role player is able to manipulate with such formats and record different information in his *CoC*. The provided framework will let the role player to define his own custom terms to describe different forensic information recorded on the *CoCs*.

C. Provenance of Information

Provenance of information is an essential ingredient of any tangible *CoC* quality. The ability to track the origin of data is a key component in building trustworthy, which is required for the admissibility of any digital evidence.

Classically, the provenance information about 'Who' created and published the data and 'How' the data is published provides the means for quality assessment. Such information can be queried and consumed to identify also the outdated information. *CoC* data source should include provenance metadata together with the forensic data. Such metadata can be exploited to give the juries data clarity about the provenance, completeness, and timeliness of the forensic information and to strength the provenance dimension for the published data.

According to the literature, different methodologies are provided by the semantic web to integrate different provenance information to the published data. Such methodologies can be classified into three main categories:

First category is using the provenance vocabularies of the semantic web [18][19][45][64]. Second category is to use Open Provenance Model (OPM) [65], and the last category is the use of Named Graph (NG) for RDF triples to add provenance metadata about each group of triples.

1) First Category : Provenance Vocabularies

A widely deployed provenance vocabularies are the Dublin Core (DC) [18], Friend of a Friend (FOAF) [19], and Simple Knowledge Organization System (SKOS) [45][64](i.e., considered as built in vocabularies on the semantic web), which contain different predicates that can provides extra information related to the published data. The objects of these predicates can be represented by URI (e.g., dereferenceable resources) or literal/terminal identifying such objects. Another provenance vocabularies provided in [17][66] describe how provenance metadata can be created and accessed on the web of data. All vocabularies presented in the semantic web can assess the quality and trustworthiness of any published data.

An example is shown in the Figure 5. In this figure, a person called Richard Cyganiak identified himself by URI http://richard.cyganiak.de/foaf.rdf#cygri and he used the *rdf:type* to specify the person class, and the FOAF vocabulary to specify his name and location. He stated that he is near to Berlin using the URI http://dbpedia.org/resource/Berlin represented in the name space dbpedia:Berlin. The latter is dereferenced and can be a subject for another RDF graph describing the City Berlin in more details: its population in which country this city is located.

Finally, the third RDF graph used the name space object of the second graph to provide what are the other cities located in Germany. All Data are expressed and enriched using different semantic web vocabularies.

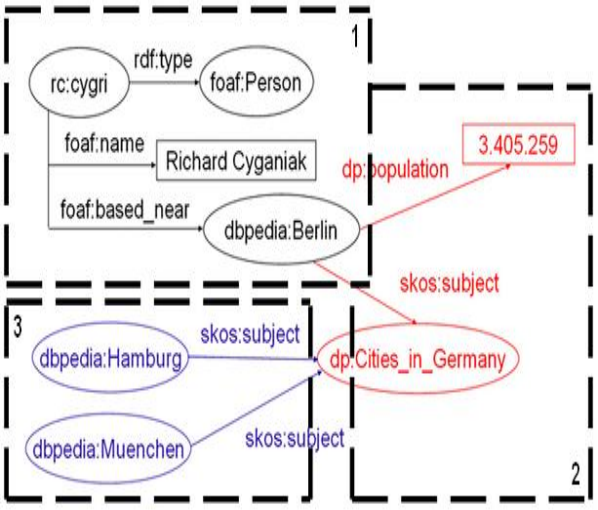


Figure 5. RDF Model with Provenance Vocabularies [67]

2) Second Category : Open Provenance Model (OPM)

Open Provenance Model (OPM) is a more expressive vocabulary that describes provenance in terms of agents, artifacts, and processes [65][68]. An extension of this work is the Open provenance model vocabulary (OPMV) provided in [65], that implements the OPM model using lightweight OWL. Open Provenance Model Vocabulary OPMV can be used also with other provenance vocabularies such as DC and FOAF.

3) Third Category : Named Graph

Whilst many authors advocate the use of semantic web technologies (i.e., vocabularies, Light weight ontologies), Caroll et al. [69] take the opposite view and proposed Named Graphs as an entity denoting a collection of triples. The idea of the named graph is to take a set of RDF triples, and considering them as one graph and assign to it a URI reference.

The NG is useful to the juries to navigate and access provenance metadata related to certain set of triples and get more description about them (e.g., LDspider [70] allows crawled data to be stored in an RDF store using the named graphs data model). As the SPARQL is widely used for querying RDF data, it can also be used in the named graph to query single or sets of named graphs. Recent work published in [23] allows publishers to add and trace provenance metadata to the elements of their datasets. This is presented through the extension of the void vocabulary into *voidp* vocabulary (i.e., lightweight provenance extension for the void vocabulary) [71]. This vocabulary considered different properties such as dataset signature, signature method, certification, and authority in order to prove the origin of a dataset and its authentication.

Simply, to illustrate how the NG can be applied in this context. If we imagine that a forensic phase (e.g., named graph of authentication, NG_{Auth}) imported from a forensic process (e.g., Kruse model) can be represented by a set of triples. Thus, the idea of the named graph will be to take this set of RDF triples (i.e., Graph) and name this graph with a URI reference.

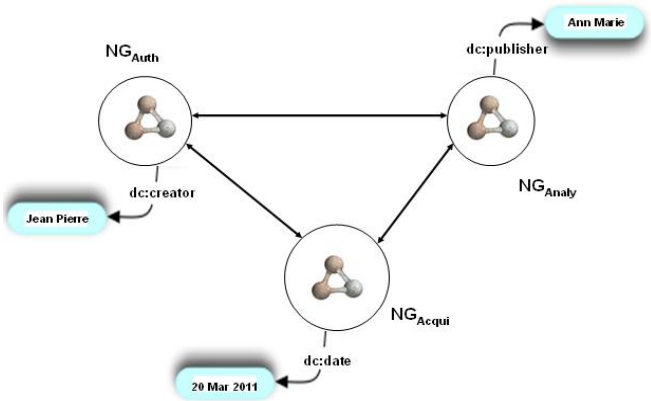


Figure 6. Named graph for Kruse Model

Figure 6 provides an abstract diagram depicting the grouping of triples and naming them to a graph with the integration of provenance metadata (e.g., DC). Each phase contains will also contain inner and outer links that relate all *CoCs* to each others.

So, the provenance metadata can be added on different levels. They can be added during the design of terms (i.e., to describe the term itself), during the publication of terms (i.e., to add more information about the data being published), or after grouping set of triples together and naming them using URI reference.

D. Public Key Infrastructure

Provenance metadata are not sufficient to ensure that the published data belong to the right players. PKI approach allows juries to ensure from the identity of role players participated in the forensics investigation.

The most related work in literature related to this paper is the one provided by Rajabi et al. in [72]. They explained theoretically how PKI is used to achieve the trustworthiness of LD and how different datasets are exchanged in a trusted way. As well, the work provided by M. Cobden et al. in [25], outlined in a vision paper, the need to have an access restriction on the LOD. Each work apart does not provide the complete picture to realize the LCD using PKI. In [72], the work explains how the PKI can be used to secure the resources of LD, but did not put the scope on how such stuffs can be implemented and applied, or how this work can bring out a new era of research related to the counter part of LOD (i.e., LCD). However, in [25], the work outlined the need of the LCD in certain domain (e.g., business and finance), but did not refer to the PKI solution, or how the LCD can be realized. Thus, this paper complements and completes the half picture of both works, by explaining how the PKI and Digital Certificates (DC) are used to restrict the access of resources in the LD cloud while keeping the resolvability of such resources, and then resulting the LCD.

This section underlines some concepts from literature related to the PKI especially the DCs; what are DCs, their purposes, their protocols, how they are created, what their types are, and how they can be exchanged. In addition, it will explain how the authors of this paper adapt PKI to LOD.

1) PKI and Digital Certificates

PKI is a combination of softwares and procedures providing a mean to create, manage, use, distribute, store, and revoke digital certificates [73][74][75][76]. PKI called Public Key because it works with a key pair: the public key and the private key

A digital certificate is a piece of information (e.g., like a passport) that provides a recognized proof of a person/entity identity. It uses the key pair managed by the PKI to exchange securely the information in order to create trustworthiness between data provider and data consumer in a network environment [77] (i.e., trustworthiness occurs when receiver ensures from the identity of the sender. As been mentioned it is known as non-repudiation).

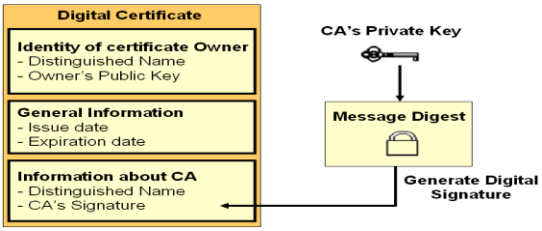


Figure 7. Digital certificate

Any certificate (see Figure 7) contains the identity of the certificate owner, such as distinguisher's name, and information about the CA (issuer of certification), such as CA's signature of that certificate, and general information about the expiration and the issue date of that certificate [78].

Digital certificate alone can never be a proof of anyone's identity. A third trusted party is needed to confirm and sign the validity and authority of each certificate and share securely the cryptographic key pair. This party is called Certification Authority (CA).

Since a CA (e.g., VeriSign Inc., Entrust Inc., Enterprise Java Bean Certificate Authority-EJBCA, etc.) relies on public trust, it will not put its reputation on the line by signing a certificate unless it is sure of its validity, the fact that makes them acceptable in the business environment.

All digital certificates provide the same level of security, whether they are created by a well-known issuer, or by unknown one. Usually, the information providers request their certificates from well-known parties when they provide services and information with large segment in society. In this paper, the authors imitate the issuer party and create CA certificate instead of buying it from well-known trusted party.

2) Purposes

A digital certificate has various security purposes and can be used to [74]:

- Allow only the authorized participant (sender/receiver) to decrypt the encrypted transmitted information (i.e., encryption).
- Verify the identity of either sender or recipient (i.e., Authentication).
- Keep the privacy of transmitted information only to the intended audience (i.e., privacy/confidentiality).
- Sign different information in order to ensure the integrity of information and confirms the identity of the signer of such information (i.e., digital signatures). Digital signatures also solve the non-repudiation problem by not allowing the sender to dispute that he was the originator of the sent message.

3) Protocols

In the field of ICT, the digital certificate is called SSL/TLS certificate because it uses two essential protocols; the SSL and the TLS [79]. The former is the short version of the secure socket layer. This protocol is used to describe a security protocol underlying a secure communication between a server and a client. After upgrading this protocol

with some encryption standards, the protocol got another acronym called TLS, which is standing for Transport Layer Security. Both protocols are based on the public key cryptography [78]. They are used to establish a secure connection over the HTTP. Classically, the HTTP establishes an unencrypted connection without using the SSL and TLS (i.e., if there is some intruder around monitoring the communication between server and client, he can come with all plain data packages of such transferred data). HTTP is then extended to HTTPS to secure the connection and encrypt all the transferred data with the SSL (i.e., HTTP + SSL/TLS = HTTPS) [80].

4) Creation Phases

The creation of a digital certificate passes by four phases (see Figure 8) using the OpenSSL tool [81]. First step, the requester (client/server/CA) generates his own pair of keys (i.e., key file), then he creates a request (i.e., req or csr format file) to the trusted party to issue for him/her a certification (i.e., crt format file). The trusted party (i.e., CA) signs the request and issues the certificate using his own private key (i.e., when the CA is the requester of the certificate, then this certificate is considered a self-signed certificate/root certificate). The created certificate is then transformed to an exportable format (i.e., p12 format) for sending it to the requester (i.e., in our context the requester is the role player).

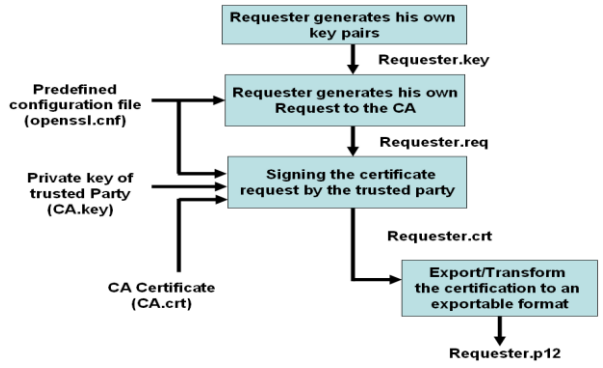


Figure 8. Procedures of creating a digital certificate using openSSL tool [1]

5) Types and Exchange

There exist three types of digital certificates. Figure 9 presents an abstract scenario where Alice and Bob want to share information over a secure connection (i.e., HTTPS).

Firstly, Alice and Bob should determine a third trusted party called the CA. The latter is responsible to issue SSL/TLS certificates for both of them in order that each can identify himself/herself to the other. CA issues two types of certificates.

- **Server certificate***:* this certificate is issued by the CA and it is used by Alice (i.e., suppose that she is the owner of the information) to identify herself to her authorized clients, like Bob. When Bob tries to access this server, he will be sure that he accessed the right server. Otherwise, Bob will not trust Alice information.
- **Client certificate**: the CA issues this certificate, and it is used by Bob (i.e., suppose he is the consumer of Alice' information) to identify himself to Alice. Alice will not allow any one to access her information unless he has a certificate known by her.
- **CA certificate:** CA also has the own certificate to sign the certificate requests received from the clients and servers. In addition, this type of certificate answers the question of how Alice and Bob ensure the identities of each others. Alice would know that Bob is the right person by verifying that his certificate is signed by the common trusted part authority (CA), as well as for Bob. Both know each others through the CA certificates.

From the definitions mentioned above, we notice that there is no distinguishable difference between the server certificate and the client certificate; both use the certificates to identify themselves to each other. However, the only difference that distinguishes both is about who is providing the information and who will go to consume it.

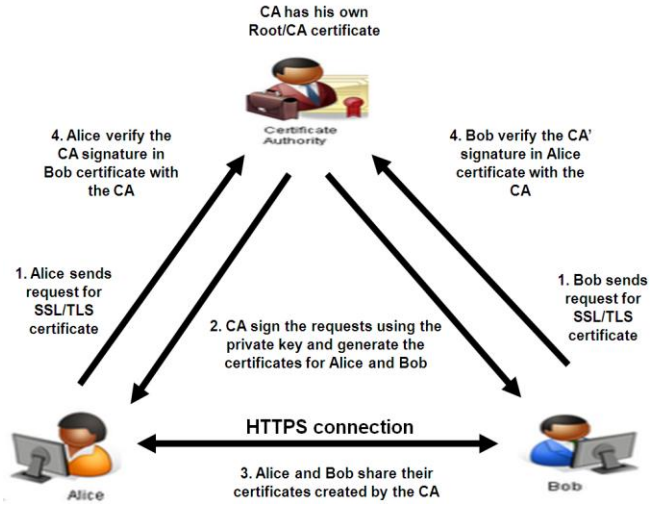


Figure 9. Sharing SSL/TLS certificates [1]

6) Adapting PKI to LOD

In this sub-section, we will discuss how digital certificates can be applied to LOD to publish and consume data on a small scale. In other words, this section describes how digital certificates are used to restrict the access of certain resources and at the same time, such resources will be resolvable to more resources.

Referring to Figure 3 of the linking open data cloud diagram, we find several data sets interrelating using outer and/or inner links. Each data set is published in a unique domain owned only by the publisher of this data set over the WWW space. Each data set contains set of URI resources that are interrelated between each others within the same data set or to an outer data set.

Now, imagine that the owner of a data set wants to publish resources using the technology stack/LDP of the LD (URI, HTTP, and RDF) and having such resources resolvable within the LOD cloud, but at the same time, he wishes to publish them in a manner that any anonymous parties on the web space cannot access them.

The idea to realize both features at the same time (i.e., resolvability and access restrictions of resources) resides in the digital certificates. The latter can be used to restrict the resolvability of resources in a one-way manner. With other words, the resources are restricted using digital certificates to be forward resolvable, but not backward resolvable unless the owner of such resources specify and list his authorized clients existing outer of his domain to access his resources. Same concepts can be applied between data sets/resources in the LOD cloud, where each data set owns a digital certificate(s). Thus, publisher of the resources can accomplish his publication task through an enhanced technology stack using a secure access protocol (i.e., HTTPS). Therefore, the current technology stack is transformed from (URI, HTTP, and RDF) to (URI, HTTPS, and RDF).

Imagining a scenario will be as follow: assuming that the publisher (server) and consumer (client) of the LD have already a common trusted party to issue their certificates. The publisher has a domain name named by an IP (i.e., for simplicity consider this IP is corresponding to a domain string name in [82]) to publish his resources in the LOD cloud. The publisher of this domain wants only someone called: 'Jean-Pierre' to consume his resources from his domain within the LOD cloud. In this case, the publisher of the data has restricted the access to his resources to a specific consumer, but he is still able to dereference his resources and resolve them to retrieve more resources outside his dataset/domain. Publisher will be also able to move back to his domain using the backward link, because he owns the server certificate for this domain. Any other anonymous party outside this domain will not be able to access the resources of [82]. If the publisher wants someone else rather than 'Jean-Pierre' accesses his resources, this person should have a client certificate signed by the same trusted party.

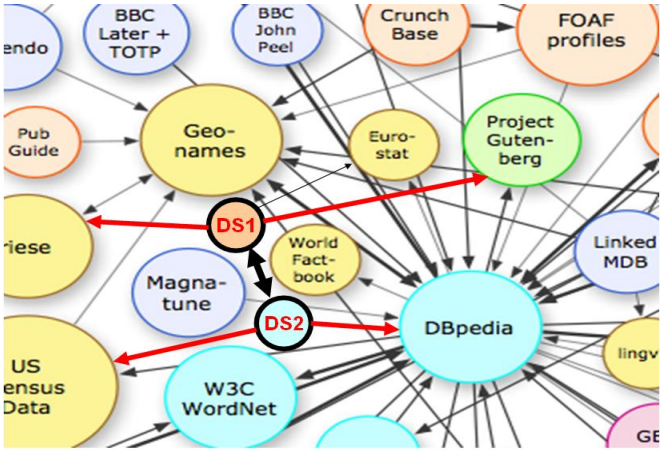


Figure 10. Client/Server certificate between two data sets

Talking in a LD manner, we can not only consider the client side as a person (i.e., as Jean-Pierre to access restricted resources), but the client side can also be a dataset or a resource within a data set that can access other resources in another data set using outer links (i.e., by moving backward to the publisher resources). In addition, another important point should be underlined; Jean-Pierre/dataset/resources can react also as a server side, if we look to the picture from the inverse direction.

Thus, Jean-Pierre/dataset/resource may have also a server certificate for his/its domain and allows the access to only people/dataset/resource that has a client certificate to his/its domain.

To illustrate this idea, Figure 3 of the LOD cloud is zoomed-in, resulting in Figure 10. Let us consider that we have two data sets DS1 and DS2 residing in two different domains. Each domain represents a data set. Both of them are interrelated between each others using inner and outer links. As well, both data sets are related with other data sets in the LOD cloud.

To elaborate the idea in terms of dataset, let consider the DS1 and DS2 can be client and server at the same time. If we look from the DS1 to DS2, we will see an outer link from DS1 to DS2 and vice-versa. DS1 is considered as a client trying to access the server DS2. Thus, DS1 will have a client certificate for its domain to identify itself to the server certificate installed in the DS2 domain. Now, let us consider if we have the contrary view; DS2 should has then a client certificate to access the server DS1 resources. However, for any other data sets around the scope of DS1 and DS2, they will not be able to resolve their resources with resources from DS1 and DS2 (i.e., at this time, DS1 and DS2 act as servers and requires client certificates from their surrounded data sets). Therefore, the resources of DS1 and DS2 have access restriction while their resources are resolvable with different resources from the LOD cloud, but the latter cannot resolve their resources from the two data sets, DS1 and DS2.

Furthermore, the certificates cannot only used on the level of datasets (i.e., including all resources), but can also be issued on the level of a specific resource within the datasets. This can be realized by issuing the certificate using one of the three URI patterns provided in Section II.

E. Pattern Consumption Applications

LD is a style of publishing data that makes it easy to interlink, discover, and consume them on the semantic web. The first way to publish LD on the web is to make URIs that identifies data items dereferenceable into RDF descriptions. Consumers can use three different patterns (i.e., in this context it will be the juries) to consume the information (i.e., the *CoC* published by role players): browsing, searching, and querying. Browsing is like the traditional web browsers that allow users to navigate between HTML pages. Same idea applied for LD, but the browsing is performed through the navigation over different resources by following RDF links and downloads them from a separate URL (e.g., RDF browsers such as Disco, Tabulator,

or OpenLink Browser). A custom semantic specialized for the juries can be easily created [83].

RDF Crawlers are developed to crawl LD from the web by following RDF links. Crawling linked LD is a search using a keyword related to the item in which juries are interested (e.g., SWSE [84] and Swoogle [85]). Juries can also perform extra search filtering using query agents. This type of searching is performed when SPARQL endpoints are installed, allowing expressive queries to be asked against the dataset. Furthermore, a void vocabulary (vocabulary of interlinked datasets) [71] contains a set of instructions that enables the discovery and usage of linked datasets through dereferenceable HTTP URIs (navigation) or SPARQL endpoints (searching), using SPARQL protocol (*void:sparqlendpoint*) or URI protocol.

In this work, we will present a framework to solve different problems related to these facets. Nowadays, the tangible CoC presenting the digital evidences and their forensic information, need to undergo a radical transformation from paper to electronic data readable, discoverable, understandable, and consumable by people and computers. This transformation helps to accommodate the evolution of digital technologies. In addition, the nature of the cyber forensic field needs a solution that unifies and represents forensic information and its formats [63] in a unified framework [14] (i.e., first and second facets). In addition, the forensic information that we want to represent needs to be shared securely on a small scale between only the role players and juries. This fact necessitates the usage of a security algorithm (i.e., PKI approach, fourth facet). Furthermore, the security of information is not enough to build trustworthiness between both parts (i.e., publisher of information represented by role players, and consumer of this information represented by juries). The admissibility of the represented information is also mandatory. The ability to track the origin of data is a key component in building trustworthy of information in order to be admissible and accepted in the court of law. Thus, Provenance of information is also required (i.e., third facet). Finally, the forensic and its provenance information needed to be available and consumed by juries through different patterns consumption applications. The latter will help them to understand and take the proper decision towards the represented information. These problems will be explained also in details in Section IV.

III. ADVANTAGES OF USING LDP FOR REPRESENTING COC

This section depicts explicitly all the advantages of using the LDP to represent the *CoC*. Knowledge *representation* has been persistent at the centre of the field of Artificial Intelligence (AI) since its founding conference in the mid 50's. Davis et al. [86] describe this concept through five distinct roles. The most important is the definition of knowledge representation as a surrogate for things. Thus, before explaining how each layer in the solution framework works, the authors decided to underline why LD is selected to represent the tangible *CoC* in cyber forensics. Thus, this section lists all the advantages and the common features of using LD to represent the *CoC* for cyber forensics:

1. *CoC* and LDP are metaphors for each others. The nature of *CoC* is characterized by interrelation/dependency of information between different phases of the forensics process. Each phase can lead to another one. This interrelation fact is the basic idea over which the LD is published, discoverable, and significantly navigated using RDF links. RDF links in LDP will not be used only to relate the different forensic phase together, but it can also assert connection between the entities described in each forensic phase. In addition, RDF typed links enable the data publisher (role player) to state explicitly the nature of connection between different entities in different and also same phases, which is not the case with the un-typed hyperlinks used in HTML.
2. LD enables links to be set between items/entities in different data sources using common data model (RDF) and web standards (HTTP, URI, and URL). As well, if the *CoC* is represented using the LDP, the items/entities in different phases can be also linked together in forensics process. This will generate a space over which different generic applications can be implemented:
   - *Browsing applications*: enable juries to view data from one phase and then follow RDF links within the data to other phases in the forensics process.
   - *Search engines*: juries can crawl the different phases of the forensics process and provide sophisticated queries.
3. LD applications that are planned to be used by juries will be able to translate any data even it is represented with unknown vocabulary. This can be realized using two methodologies. First, by making the URIs that identify vocabulary terms dereferenceable (i.e., it means that HTTP clients can look up the URI using the HTTP protocol and retrieve a description of the resource that is identified by the URI) so that the client applications can look up the terms, which are defined using RDFS and OWL. Secondly, publishing mappings between terms from different vocabularies in the form of RDF links. Therefore, for any new term definition, the consumption applications are able to provide and retrieve for the juries extra information describing the provided data.
4. Nowadays, RDFS [15] and OWL [38] are partially adopted on the web of data. Both are used to provide vocabularies for describing conceptual models in terms of classes and their properties (definition of proprietary terms). RDFS vocabularies consist of class *rdfs:class* and property *rdf:property* definitions, which allow the subsumption relationships between terms. This option is useful for juries to infer more information from the data in hand using different reasoning engines. For

example, RDFS uses a set of relational primitives (e.g., *rdfs:subclassof*, *rdfs:subpropertyof*, *rdfs:domain*, and *rdfs:range* that can be used to define rules that allow additional information to be inferred from RDF graphs). Also, OWL extends the expressivity of RDFS with additional modeling primitives that provide mapping between property terms and class terms, at the level of equivalency or inversion (e.g., *owl: equivalentProperty*, *owl:equivalentClass*, *owl:inverseof*).

RDFS and OWL are not yet fully adopted on LDP, but soon the full adaptation will be achieved [87][88][89]. This will be a great advantage to add more property and class terms to the semantic dimension of the LD, and therefore, provide useful and descriptive information.

5. Representing *CoC* data using LDP will be enriched with different vocabularies such as Dublin Core (DC) [18], Friend of a Friend (FOAF) [19], and Semantic Web Publishing (SWP). In addition, vocabulary links is one type of RDF links that can be used to point from data to the definitions of the vocabulary terms, which are used to represent the data, as well as from these definitions of related terms into other vocabularies. This mixture is called schema in the LD; it is a mixture of distinct terms from different vocabularies to publish the data in question. This mixture may include terms from widely used vocabularies as well as proprietary terms. Thus, we can have several vocabulary terms to represent the forensics data and make it self descriptive (i.e., using the two methodologies mentioned in point 3) and enable LD applications to integrate the data across vocabularies and enrich the data being published.
6. Juries need to avoid heterogeneity and contradictions about the information, which are provided to them in the court in order to take the proper decision. LD tries to avoid heterogeneity by advocating the reuse of terms from widely deployed vocabularies (same agreement of ontology). LDP is then useful to represent this type of information.
7. As mentioned in point 1, a forensics process contains several phases, which are dependent and related to each others. Each entity is identified by a URI namespace to which it belongs. An entity appearing in a phase may be the same entity in another phase. The result is multiple URIs identifying the same entity. These URIs are called URI aliases. In this case, LD rely on setting RDF links between URI aliases using the *owl#sameas* that connect these URIs to refer to the same entity. The advantages of this option in *CoC* representation are:
   - *Social function*: investigation process is a common task between different players. The descriptions of the same resource provided by different players allow different views and opinions to be expressed.
   - *Traceability*: using different URIs for the same entity allows juries that use the *CoC* published data to know what a particular player in the investigation process has to say about a specific entity of the case in hand.

   Same thing occurs not only at the level of URI but also at the level of terms. Players of the forensics process may discover at a later point that a property vocabulary contains the same term as the built in one. Players could relate both terms, stating that both terms actually refer to the same concept using the OWL (*owl: equivalentClass*, *owl: equivalentProperty)* and RDFS vocabularies (*rdfs: subclassOf*, *rdfs:subProperyOf*).
8. Provenance metadata can also be published and consumed on the web of data [66]. Such metadata provide also an answer to six questions, but at the level of the data origin (i.e., see Section I for provenance questions). These vocabularies can be used concurrently with the forensics data, to describe their provenance and complement the missing answers related to the forensic investigation.

IV. Research Problems And Methodologies

As been mentioned, the *CoC* is a testimony document that records all information related to the evidences (digital/physical) in order to ensure that they are not altered throughout the forensic investigation. Failure to record enough information related to the evidence may lead to its exclusion from the legal proceedings.

Nevertheless, the existences of many works for the *CoC* in CF there are still several issues preventing the role players to securely record, describe, and manage the results of their forensic investigation. In addition, these problems complicate the task for juries to consume and understand the digital evidences and take the proper decision about the provided information. Some problems may be resolved in the literature using another way or using classical (non-technological) methods. This section will summarize explicitly how the novel CF-*CoC* framework uses new existing technologies to solve such problems.

A. First problem : Accomodation with technology evolution

As mentioned in Sections I and II, cyber forensic is a daily growing field that requires the accommodation on the continuous changes of digital technologies (i.e., concurrency with the knowledge management). Each forensic process is associated a tangible *CoC* document that need to undergo a radical transformation from paper to machine-readable format to accommodate this continuous evolution. The LD has widely established de-facto standards (RDF, SPARQL) for sharing and interlinking of data on the semantic web.

In addition, the forensic information resulted from the forensic tools need to be interoperable with the represented *CoC*s in order to obtain a complete picture about the accomplished investigation process. AFF4 [61][63] is an open format for the storage and processing of digital evidence. Its design adopts a scheme of globally unique identifiers (URN) for identifying and referring to all

evidence [63]. The great advantage of this format is representing different forensic metadata in the form of RDF triple (subject, predicate, and value). The subject is the URN of the object the statement is made about, and the predicate (e.g., datelogin, datelogout, evidenceid, affiliation, etc.) can be any arbitrary attribute, which can be used to store any object in the AFF4 universe. Representing such formats in the same unified framework (RDF), facilitates as well the consumption of all information resulted from the forensic tools.

The CF-*CoC* will use the semantic web as a fertile land to create interlinking *e-CoCs* readable and consumable by the machine and the forensic information resulted from a forensic tool can be interoperable with these interlinking *e-CoCs (layer 1,2,3, and 5)*

B. *Second problem : security of represented CoC*

Second problem concerns the security of the *CoC* documents. Usually, the *CoC* documents must be affixed securely when they are transported from one place to another. This is achieved using a very classical way: seal them in plastic bags (i.e., together with physical evidence if there is any, such as hard disk, USB, cables, etc.), label them, and sign them into a locked evidence room with the evidences themselves to ensure their integrity. The *e-CoCs* need also to be secured since their publication by the role players until their consumption by the juries. LDP are used to publicly publish the data on the web and need to be adapted with some access and license restrictions.

The CF-*CoC* will use PKI to securely publish and consume the data in a small scale between the role players and juries *(layer 6)*.

C. *Third problem : Build trustworthiness between role players and juries*

The problem is not only to represent the knowledge of the tangible *CoC* in order to solve the issues mentioned above, but also to express information about where the *CoC* information came from. Juries can find the answers to their questions on the *CoC*, but they need also to know the provenance and origins of those answers. As been mentioned, provenance of information is crucial to guarantee the trustworthiness and confidence of the information provided. Provenance information is responsible to answer questions about the origin of answers (i.e., what information sources were used, when they were updated, how reliable the source was).

The CF-*CoC* will use provenance metadata imported from different vocabularies of the semantic web. Such vocabularies can be useful to answer the questions about the origin of the *CoC* data. Providing answers to such questions make the *e-CoC* admissible to the court of law *(layer 4)*.

V. CF-*CoC* Framework And System

The CF-*CoC* framework presented in Figure 1 is constituted of several layers. Each layer is responsible to perform certain task. The order in which the layers are placed is just to provide a conceptual diagram and explains the different tasks needed to convert a tangible *CoC* into electronic data. The number assigned to each layer, it is just for numeration. For example, the PKI layer is numerated by number six, and it is placed as last layer. This does not mean that it will be used as a last task. However, it can be used along several tasks (i.e., before defining terms, during publishing of terms, or during consumption). It is just placed at the top to globalize that it can be applied to any of their antecedent layers. Another example is the provenance of information layer, this layer can be applied to the term being designed, or to the terms being published. As we mentioned in Section II: the provenance of information may be used to describe the terms during their design, during their publication, or after publishing set of triples describing certain forensic phase.

Thus, the provenance layer describes the addition of different provenance metadata to the forensic information being published. Juries can then query and consume this information and its metadata from the consumption layer. The latter provides different consumption applications to the juries in the court of law.

This section describes in details how each layer can be built and implemented. Different modules are implemented in the CF-*CoC* system (see Figure 11), and each module contains different tasks.

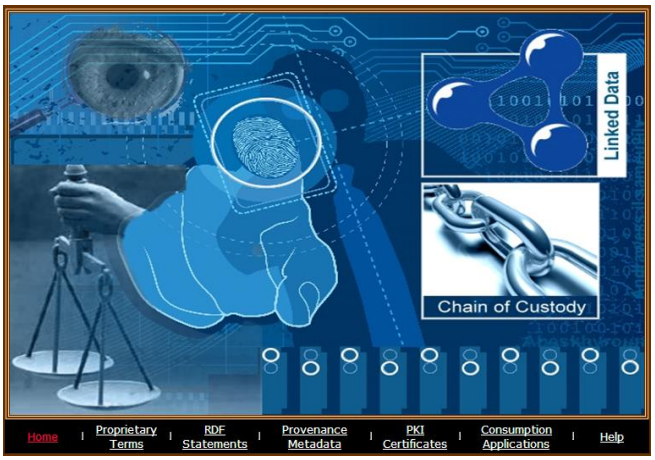


Figure 11. CF-*CoC* System

First module dedicated to create proprietary terms. This module is used to identify and design new custom terms to describe forensic information. A module for creating set of RDF statements. This module is used to publish different RDF triples by using different proprietary terms (i.e., custom terms created by publisher/role players) and build in terms (i.e., terms defined by well-known vocabularies of the semantic web). Different provenance metadata can be added along the tasks of both modules using the provenance metadata module. A PKI certificates module is also integrated on the system to create different types of

certificates to bend the designed and published terms and triples on a small scale to be shared only between juries and role players. Finally, juries consume such represented information using the consumption module.

A. Work Environment

The CF-*CoC* framework is implemented using Php and easyRDF [90], Graphiz tool [91], and its graph objects are used within the easyRDF to produce and draw different RDF models. In addition, the operating system used in this experimentation is Windows XP, accompanying with the Internet Information Services (IIS) [92] and the Openssl tool [93]. IIS simulate the machine as a server, and the OpenSSL tool is used to create the digital certificates.

B. CF-CoC Terms Definitions (Layer 1 and 2)

As shown in Figure 4, the first step is to create the ontology corresponding to a forensic phase. This ontology will contain all the forensic terms describing the different tasks of acquisition phase. In the tasks of creating ontologies, proprietary terms, and publication of data using such terms, we assume that the publishers (i.e., role players of a forensic process) own background knowledge of how to create ontologies and publish RDF data. Each role player will define his own terms from his point of view.

1) Creation of Ontology (Vocabulary):

The main objective of creating ontology objects is to create proprietary terms. Ontology object in the LD acts as a container for creating custom terms. The role players are responsible to create such objects. In LD, it is sufficient to create the ontology object and add provenance information to it (i.e., publisher name, date of creation, label, and comments). After creating ontologies, the role player appends and creates proprietary terms to his ontology(ies). Other role players can share these custom terms to publish their own data. Thus, ontologies can be reused and shared by any role player, and each role player has the liberty to use an existing ontology describing certain forensic task, or create his new ontology to describe the same forensic task. In LD, there is no negative effect to create more than one ontology describing the same task, because by creating more ontologies, we have to define more terms describing these ontologies (i.e., corresponding to forensic phase). If there exist any redundant term, a mapping process can be performed to align such ontologies. Therefore, in the LD creating ontologies is an intellectual and subjective task not as the semantic web to create full and detailed common ontologies. By time, system will contain different ontologies describing different forensic phases, created by different role players that have different point of views.

The task of creating ontologies is about to create the ontology object of the acquisition phase (see Figure 4). The domain name field is required to mint the ontology to a unique domain name owned by the publisher (aspect 2). The second field is about the selection of role player certificate [1]. In addition, the value type of the role player can be a resource or a literal. Next fields are the ontology name and its label description.

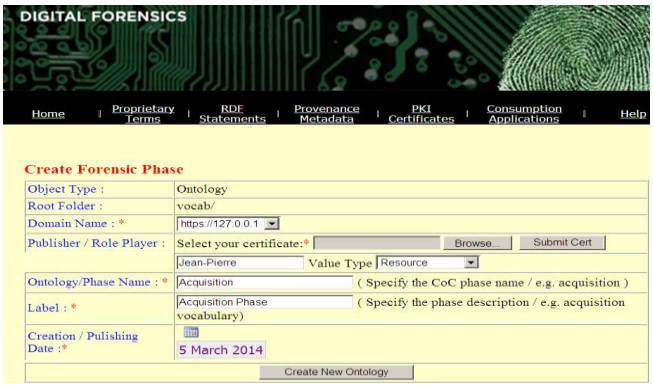


Figure 12. Creation of Acquisition Ontology

Last field is the publication date of the acquisition ontology (see Figure 12).

After completing this form, the acquisition ontology is generated by using the *Graphiz* module [91] (see Figure 13).

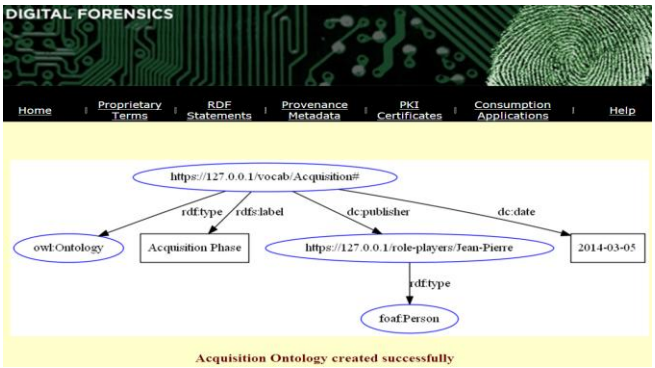


Figure 13. Creation of Acquisition Ontology

After creating the acquisition ontology, the role player proceeds with the next module to create terms and append them to this new ontology object.

2) Creation of new terms:

This task relates to four essential fields. The first field is the term name. The second field is selecting ontology to append the new proprietary term. The third field specifies the category/forensic task (see Figure 14). In our case, the category could be one of the three tasks provided in Section II (preservation, recovery, or copy). In this field, the user may select 'New' to create a new category or select 'Existing' to import an existing category, defined in another vocabulary (ontology) created by another role player (i.e., two different forensic phase may have a common category/task). Last field is the selection of term type (i.e., a term can be a property or a class).

As an example, consider the following tangible *CoC*:

"The name of the first responder in the acquisition phase is Jean-Pierre. He is the role player of this phase, and he preserved the state of the digital media, PDA device, which has the SN: 0G-4023-32-362. The date he did this task is 5 March 2014".

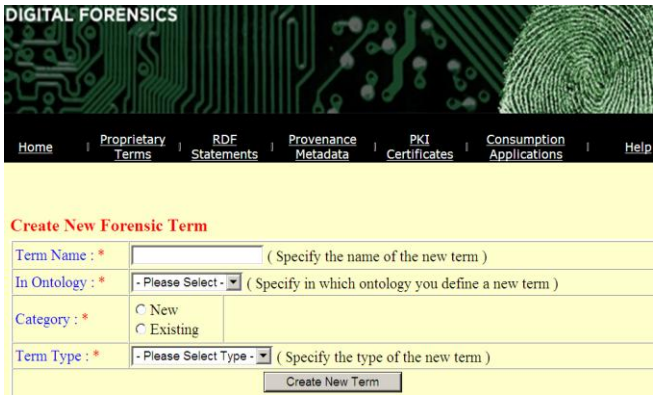


Figure 14. Creation of a New Term

The first step to create an *e-CoC* from this tangible *CoC* is to identify the terms (see Table V) (i.e., as we mentioned in Section II.A.4, identifying proprietary terms are subjective task and may differ from one creator to another).

TABLE V. PROPRIETARY TERMS OF PRESERVATION TASK

	Term name	Type
T- Box	First_responder	Class
	Role_player	Class
	Acquisition	Ontology
	Digital_media	Class
	preserve	Property
	preservedby	Property
	SN	Property
A- Box	Jean-Pierre	Subject/Object
	PDA-device	Subject/Object
	0G-4023-32-362	Object

This case study contains T-Box and A-Box information. Terms of T-Box are of type class and property. The *Role_player* term is a class that can be defined as a subclass from the class *Person* in the *FOAF* (friend of a friend) ontology (see Figure 15) [16][19]. This term will belong to a forensic task called *Preservation* task.

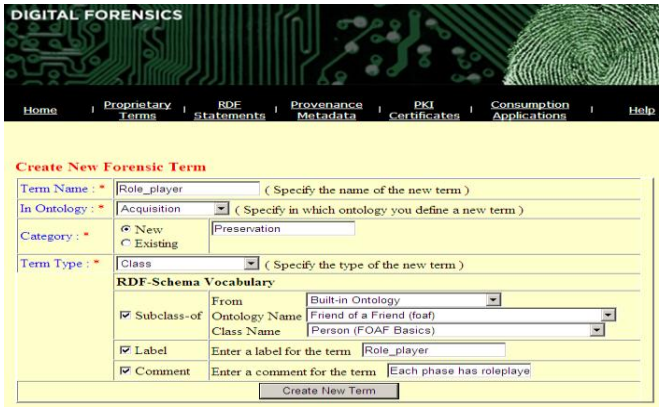


Figure 15. Creation of the Role_player Class

The *First_responder* term is a class that can be an instance of the *Role_player* class. Now, the *Preservation* category will be found under the 'Existing' category. Finally, the *Digital_media* is a subclass of *owl:Thing* (see Figure 16).

Now, the property terms (*owl:objectProperty*) will be defined. The *domain* and *range* of the term *preservedby* are defined to be *Digital_media* and *First_responder* class, respectively. This property term is defined to be a sub-property from *foaf:made* property (see Figure 17).

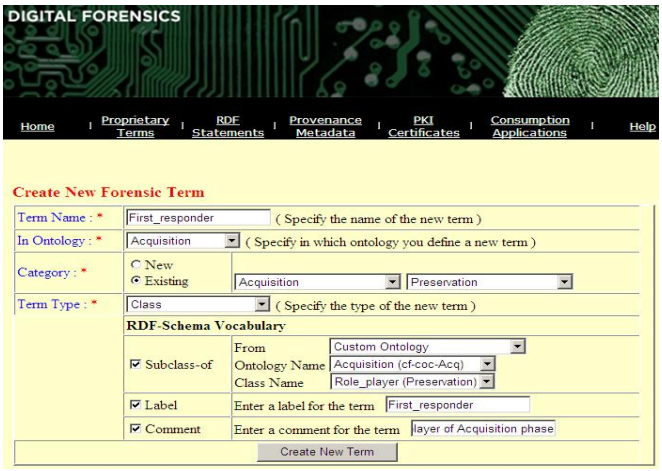


Figure 16. Creation of the First_responder Class

The *preserve* property is the inverse of *preservedby* property. Thus, the *domain* and *range* of the former will be also the inverse, *First_responder* and *Digital_media* respectively. Simply, if a digital media is preserved by a first responder, then this means that the first responder preserved the digital media. The last property is *SN*: the serial number of a device is an inverse functional property, because each serial number identifies one and only one subject (see Table III).

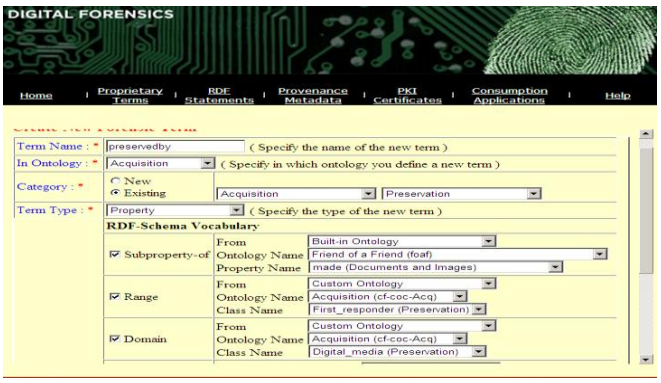


Figure 17. Creation of the preservedby Property

After creating all terms, the role player can generate the acquisition ontology with all the property and class terms of the preservation forensic task (Figure 18).

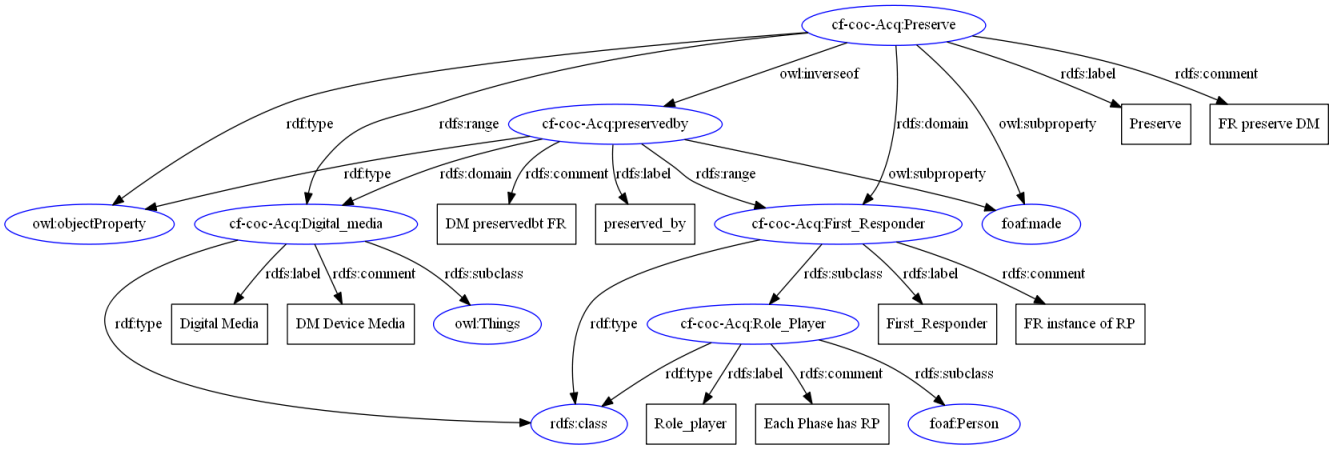


Figure 18. T-Box Ontology of Forensic Preservation Task

The T-Box terms created will be used to publish data. Therefore, they will describe the A-Box data. The latter is the *e-CoC* that will be consumed later by jury in the court using different pattern consumption (browsing [94] [95], querying [96], or searching [97]). Now, the data can be described and published using the terms defined in the T-Box and by using the third layer of the framework, the user can publish different triples (i.e., using different vocabularies of the semantic web) and by the support of proprietary vocabularies defined by the role player.

C. Publications of RDF Statements (Layer 3)

This layer is straightforward. All custom terms that have been defined in proprietary terms module (T-Box) can be used to publish and describe the *CoC* in form of RDF triples. Not only custom terms are used to publish RDF statements, but also the terms from the well know vocabularies can be used to publish such RDF statements.

The main tasks in this module are the publication of terms and mapping between them. Publication of terms is about selecting the subject, predicate (property), and object. For mapping between terms, different constructors from OWL vocabulary can be used such as *equivalentProperty*, *equivalentClass*, and *sameas* (see Table III).

The main axis over which the RDF statement is constructed is the property slot of the triple. This slot is essential to publish any RDF statement, because on its left (subject) the domain of term is defined and on its right (object) the range of the term is defined (see Table II), and then in turn, the classes and subclasses are defined. Property term (predicate) is considered as the initial node of any T-Box. From this starting node, all leaves (non-terminal) are expanded until reach the literals are reached (terminal leaves).

For example, the property term 'preserve' defined in the T-Box, its domain is First_responder class (Subject) and its range is Digital_media class (object). Thus, any literal given by the publisher in the subject slot of RDF triple will be of type First_responder, which is a subclass of Role_Player, which is a subclass of class Person (i.e., defined in the foaf vocabulary); see Figure 19.

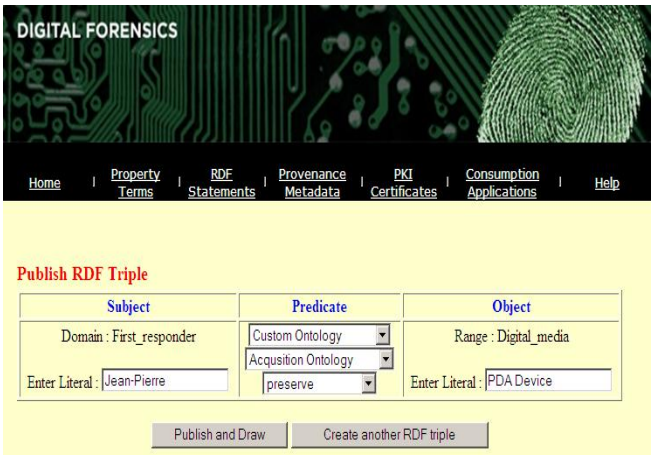


Figure 19. Publish RDF Triple

The second main task of this module is mapping between terms. The predicate slot of this triple will be one of three constructors mentioned above (see Table III), the subject and object are terms of type class or property, in order to map term class to term class or term property to term property. An example to map two different terms are those of type First_responder and Role_Player. In some cases, the role player is the generic term used to any forensic process, and in other cases the exact player is identified by its role (i.e., the role player is assigned to the acquisition phase and at the same time the player of this phase is called the First_responder). Thus, a term of type First_responder can be equivalentClass to another term of type Role_Player. Describing a term with different point of

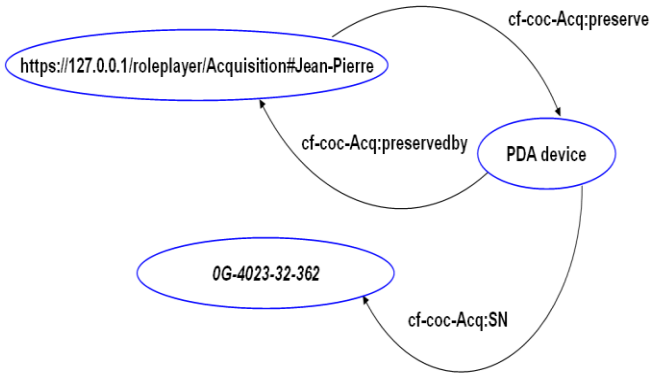


Figure 20. A-Box Ontology of Forensic Preservation Task (*e-CoC*)

views ameliorates the result of any reasoning engine (e.g., primer) [98].

Figure 20 shows the *e-CoC* (A-Box) of the forensic preservation task. This generated ontology does not answer all the question of *CoC*. It answers only the Who: Jean-Pierre, What: PDA device, and When: publication date of ontology. In order to have the answers to other questions, more terms need to be determined and defined. In this figure, the *cf-coc-Acq* is the prefix namespace of the acquisition ontology: *Jean-Pierre* is an instance from the *First_responder* class (i.e., which is an instance of the *Role_player* class), PDA device is an instance of *Digital_media* (i.e., which is instance from *Things* class), and *presevedby* is the inverse property of preserve property. *SN* is a functional property where its domain is the *PDA device* and its range is the *0G4023-32-362* (i.e., which is an instance of the *Literal* class). In addition, the forensic information resulted by the foresnic tool (e.g., AFF4) can also be represented in the *CF-CoC* framework using the same steps mentioned in Sections V.B and V.C.

D. Provenance of Information (Layer 4)

As we mentioned in Section II.C, the provenance metadata can be added to the terns during their design or to set of RDF triples. The framework CF-CoC use the named graph method to add provenance metadata to set of triples by naming them using URI.

Role players are responsible to add different provenance metadata to describe their forensic information. Each role player is responsible to provide complete and correct information about the origin and contents of his CoC(s) in order to be admissible in the court of law. In this context, data is shared on a small scale (i.e., LCD) between role players and juries. Identities of role players have been validated before the investigation process through exchanging of digital certificates. Thus, adding provenance metadata manually to the forensic information being published does not affect the creditability of these metadata. However, this is not the same case of sharing data on opened scale (i.e., LOD), where public data needed to be tracked and verified in order to ensure its creditability.

An example of metadata added to the level of terms presented in Figure 13, where the DC vocabulary is used to answer when the ontology is published and who published it. Provenance metadata can be attached during the phase of T-Box and A-Box.

Figure 20 is a good example to add provenance metadata using the named graph method. This figure represents the task of state preservation in the cyber forensic acquisition phase. Figure 6 provides abstract models for the named graph. The NG_{acqui} is the named graph of acquisition phase, which contains three tasks. One of them is the preservation task provided in Figure 20.

Figure 21 depicts how provenance metadata is added to a named graph. The CF-*CoC* assigns automatically the URL address to each ontology by adding a suffix NG to the ontology URL. For example, if the URL of acquisition ontology is https://127.0.0.1/Acquisition.rdf then it will be https://127.0.0.1/AcquisitionNG.rdf. In the same screen, the CF-*CoC* requires to select the ontology from which we will select the desired property from different provenance vocabulary (e.g., DC, FOAF, etc.).

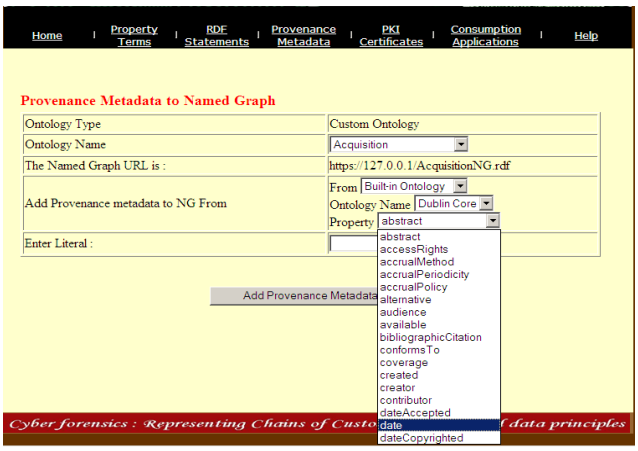


Figure 21. Add Provenance Metadata to Named Graph

As shown in Figure 21, the user selects the vocabulary/ontology in order to select the desired property, After selecting the property the user enters the literal representing the object. Thus, this is also considered as an RDF triple, where the subject is the URL address of the named graph of the acquisition phase, the predicate is the property of provenance vocabulary, and the object is the literal value.

E. PKI (Layer 6)

This section explains how to create the digital certificate using the four procedures mentioned above (see Figure 8).

Before creating the server and client certificate, a CA certificate will be created to sign both client and server requests (i.e., in this scenario, we will create manually a CA instead to buy it from a well-repudiated CA). Usually, the CA certificate is provided by a well-known CA provider (e.g., VeriSign Inc, Entrust Inc, etc.). In this scenario, a CA self-signed certificate is manually created.

The next part will present the set of theoretical procedure that juries and role players use together to share the digital certificates. This part will be followed by a detailed explanation of how these theoretical parts can be implemented and realized:

1. Juries send a list of players who are supposed to work on the current cyber crime case. Sending this list to the CA controls the data access to only these players. This prevents the disclosure (keeps the confidentiality) of data to unauthorized people.
2. The role player generates a public-private key pair ({KU-P, KR-P}), where P is all information identifying the player, R is private, and U is public. The player stores the private key in a secure storage to keep its integrity and confidentiality, and then sends the public key KU-P to the CA.
3. The player's public key and its identifying information P are signed by the authority using its ({KR-CA}) private key. The resulting data structure is back to the role player. R-CA {P, KU-P} is called the public key certificate of the role player, and the authority is called a public key certification authority (i.e., symbols outside brackets mean the signature of the data structure).
4. Juries obtain the authority's public key {KU-CA}.
5. Each player creating a *CoC* must authenticate himself to juries by signing his RDF graph G using his private key R-P{G} (i.e., all triples describing a phase are assembled in one graph called G). Later, before the court session, each player sends the certification R-CA {P, KU-P} to juries accompanied with the signed graph R-P{G}.

The next part will explain how such procedures can be implemented using the SSL tool:

1) Self-Signed Certificate:

Before starting, the CA key is generated, *RootCA.key* of length 2048 bits (2 bytes).

openssl genrsa –out RootCA.key 2048

The *RootCA.key* is then used to generate the certificate request *RootCA.csr* by providing the country name (i.e., C=CA), the organization name (i.e., O=Cyber Forensics Institution), and the common name of the certificate (i.e., CN=CF-CA) (see Figure 22).

openssl req -new -key RootCA.key -out RootCA.csr -config openssl.cnf -subj "/C=CA/O=Cyber Forensics Institution/CN=CF-CA/"

After generating the RootCA.csr, the request is signed using the RootCA.key to generate the requested certificate (*crt* format, *RootCA.crt*), but in this type of certificate, the CA itself will sign the certificate, that's why it is called self-signed certificate:

openssl req -x509 -days 365 –in RootCA.csr -out RootCA.crt -key RootCA.key -config opensslCA.cnf -extensions v3_ca

Finally, the exportable format p12 is generated to transform the *RootCA.crt* into an exportable format *RootCA.p12*

openssl pkcs12 -export -in RootCA.crt -inkey RootCA.key -certfile RootCA.crt -out RootCA.p12

Figure 22. CA self signed certificate

2) Server Certificate:

The server certificate is created for two goals: it lets the role player ensures the identity of the server, as well it is used to check for the client certificate (see Figure 23).

Assume that the server IP is corresponding to the server in [82]. This certificate will be issued for the juries to install it on their server. This server will host the *CF-CoC application*, which will be used by the role player. Thus, the CA will issue and sign a certificate for this IP name.

First, the *Server.key* is generated using the following command:

openssl genrsa -out Server.key 2048

The *Server.key* is then used to generate the certificate request *Server.csr* by providing the country name (i.e., C=CA), the organization name (i.e., O=Cyber Forensics Institution), and the common name of the certificate (i.e., CN=192.168.2.12).

openssl req -new -key Server.key -out Server.csr -config openssl.cnf -subj "/C=CA/O=Cyber Forensics Institution/CN=192.168.2.12/"

After generating the *Server.csr*, the request is signed using the CA certificate *RootCA.crt* and the key *RootCA.key* to generate the requested certificate (i.e., *Server.crt*).

openssl ca -days 365 -in server.csr -cert RootCA.crt –out Server.crt -keyfile RootCA.key -config opensslserver.cnf -extensions server

Because the server certificate is signed by the CA, the *openssl* command uses a build in parameter called '*ca*', to declare that the server certificate will be signed by the CA using its key (*RootCA.key*).

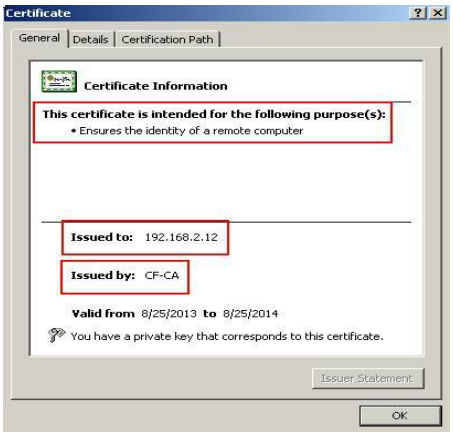


Figure 23. Server digital certificate

3) Client Certificate:

The role player authenticates himself to the server through the client certificate. Without this certificate, the role player will not be able to access *CF-CoC* application to construct different ontologies for each forensic phase and publish different resources (see Figure 24).

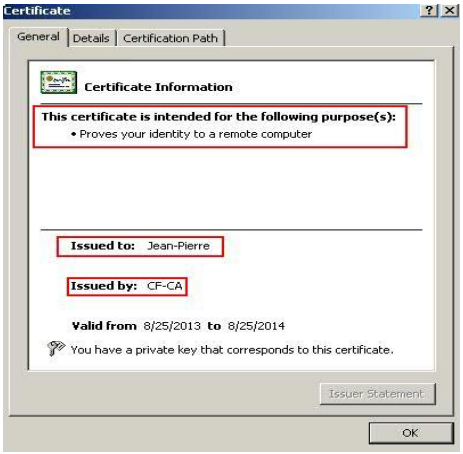


Figure 24. Client digital certificate

First, the *Client.key* is generated using the following commands:

openssl genrsa –out Client.key 2048

The *Client.key* is then used to generate the certificate request *Client.csr* by providing the country name (i.e., C=CA), the organization name (i.e., O=Cyber Forensics Institution), and the common name of the certificate (i.e., CN=Jean-Pierre).

openssl req -new -key Client.key -out Client.csr -config openssl.cnf -subj "/C=CA/O=Cyber Forensics Institution/CN=Jean-Pierre/"

After generating the *Client.csr*, the request is signed using the CA certificate (*RootCA.crt*) and key (*RootCA.key*) to generate the requested certificate (i.e., *Server.crt*).
openssl ca -days 365 -in Client.csr -cert RootCA.crt -out client.crt -keyfile RootCA.key -config opensslclient.cnf -extensions client

As shown in Figures 22, 23, and 24, we noticed that each certificate has its own purpose(s). Purpose(s) of a certificate depends on its type. The type of certificate is defined using the *-extension* in the creation of *crt* certificate. The *–extension* parameter calls the proper module for each certificate type. For example, it calls the *opensslCA.cnf*, *opensslServer.cnf*, and *opensslClient.cnf* for the CA, server, and client certificates, respectively. However, the *openssl.cnf* contains general configuration of all types of certificates.

4) Installation of Digital Certificates

Before installing the certificate, the CA sends to the jury and the role player their own certificates. Jury installs his certificate on his server and role player installs his certificate on his browser.

4.1) Self-Signed Certificate:

After creating the CA certificate, the CA sends to the server and client his certificate (i.e., p12 format without the private key of the CA certificate). By clicking on the p12 file (i.e., exportable format), a wizard will be launched to install the CA certificate in the trusted root folder of the current browsers for both server and client. By firstly installing this certificate on the server and client machines, their browsers will automatically identify the issuer of the client and server certificates.

4.2) Server Certificate:

The CA sends the server certificate to the jury. The latter then starts the installation of the server certificate. Installation of server certificates on Windows XP passes by two phases:

- Running the Microsoft Management Console and follow the steps in [99].
- Installing server certificates using the steps mentioned in [100].

4.3) Client Certificate:

Installing the client certificate is the same as the CA certificate, but at this time, the wizard installs the certificate in the client/ Personal folder of the browser.

5) Experimentation

This section shows how the scenario is enrolled after the role player and jury install their certificates:

- The client accesses the site by typing the URL of the server 192.168.2.12

- Because the remote server (i.e., where the *CF-CoC* web application is hosted) owns a server certificate, it requires then that his clients also owns a client certificate owned by the same trusted party (In this case, the CF-*CA*), otherwise the browser responded with a blank page (see Figure 25).

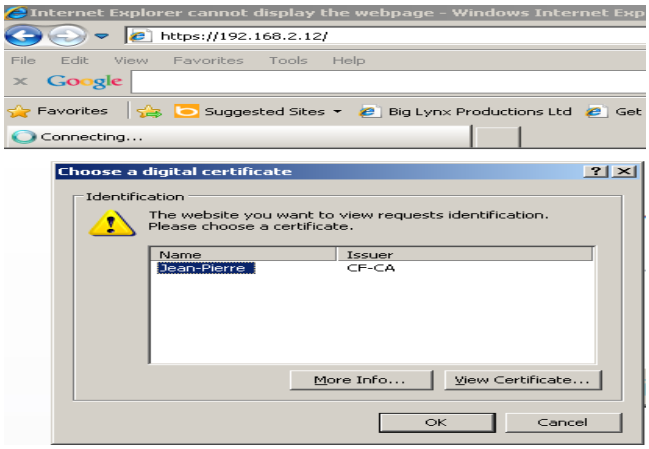


Figure 25. Server requires Client Digital certificate

- Once the server identifies the client certificate, it redirects the client to CF-*CoC* web application (see Figure 26).

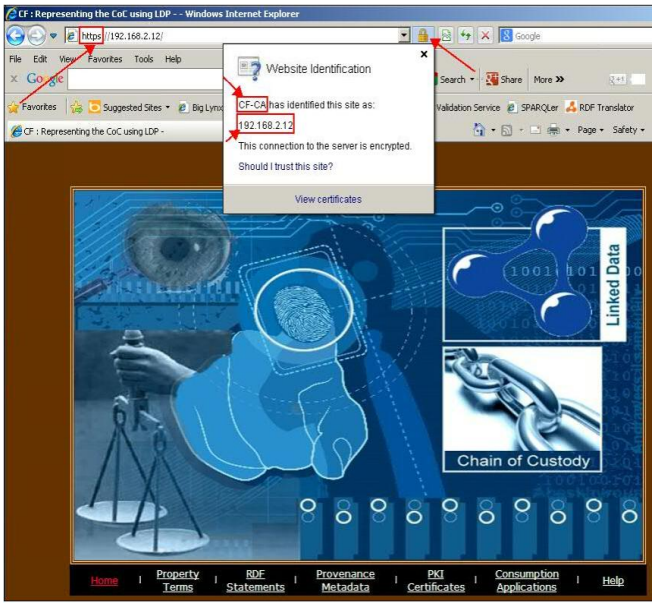


Figure 26. Redirection to the Restricted Resources

- Once the role player accesses the application, he starts to publish the ontologies and creates terms describing the forensic phase in hand (see the term definition module, Section V.B).

As we see in Figure 26, the server certificate is installed and shown in the top of the screen as a yellow lock. By clicking on the lock, it will show who issued the certificate (i.e., CA) for this page and to whom it was issued.

Once the role player finishes the publication task, the resources will be available to jury for consumption, as he owns a server certificate of the server, which allows him to view and access such resources published on his server. Resource as Jean-Pierre (see Figure 13) will be resolvable to more extra resources in the same domain [89] or to external domain. However, Jean-Pierre will not be accessible from external resources outer the former domain.

A certificate can not be created only for resources on the server but it can be issued for a specific resource on a server. For example, if we imagine that we have a resource 'x' in DS1, then the field of the certificate called 'issued to' (see Figure 23) will be assigned the complete URL of the resource 'x' (e.g., *CN=192.168.2.12/resources/x).*

F. Pattern Consumption Application (Layer 5)

After defining and publishing the foresnic information, juries can consume such resources uisng different pattern methods, whether by browsing the resources and navigate between different resoures, crawling using certain keyword, or by query the RDF triples.

The framework implemented some of them and imported the others. For example, the search and crawl was implemented, SPARQL endpoint is installed.

1) Browsing and crawling of resources

There exist many applications to browse and crawl RDF statements. All of them may have different consumption interfaces, but they are all common in the concept of how browsing and crawling are performed. In the CF-*CoC* framework, both types of consumption are simply presented by querying the RDF database and by standing on the defereneceable option to navigate between different RDF resources (see Figure 27).

Figure 27. Crawling and Browsing Consumption

Figure 27 shows the consumption screen of RDF statements. Juries can crawl a specific resources by selecting the first option and enter as keyword the required resources (e.g., Jean-Pierre), or through selecting the second option, search by forensic phase. In fact, the forensic phase

appearing to the juries is the same as the ontology terminology. As mentioned before (see Figure 4), each forensic phase is corresponding to an ontology, and each category corresponding to a forensic task in this forensic phase. For example, the preservation task is a category that contains different terms, the preserve verb (the task itself), what is the subject of this task, and who can perform this task, and what are the different ancestors of the term subject, predicate and objects.

If the user selects search by resource, he can enter a resource name to extract more information about such resource. The results of this type of search can lead to browse and dereference more related resources. For example, if the jury search for a resource called Jean-Pierre, its result will be: he was the creator of the acquisition phase. Jury in this case may get another deferenceable URL of the acquisition ontology, or another forensic task related to this forensic phase. This fact allows juries to navigate and discover more resources and retrieve more information related to all tasks performed by Jean-Pierre (see Figure 28).

Figure 28. Crawling using Jean-Pierre Resource

If the user selects the second option, he can select different ontologies that have been published by the role players, and he can select a specific task from selected ontology.

2) SPARQL Query

The SPARQL Language is the query language of the RDF. The SPARQL endpoint is installed on the local machine where the CF-*CoC* application resides.

Juries will not only able to query RDF triples, but also the provenance metadata associated to these triples. An example of how SPARQL queries the named graph and retrieve the provenance metadata such as publisher name and publishing date, is mentioned below:

```
PREFIX dc: <http://purl.org/dc/elements/1.1/>
SELECT  ?Publishername ?NamedGraph ?publicationdate
FROM NAMED <https://127.0.0.1/authenticationNG.rdf>
WHERE
{
        ?NamedGraph dc:publisher ?Publishername .
        GRAPH ?NamedGraph dc:date ?publicationdate }
}
```

Another example to query all RDF triples that are containing predicate foaf:name, and cf-coc:preserve is shown below:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX cf-coc: <http://www.127.0.0.1/vocab/acquisition#>
 SELECT * from <http://127.0.0.1/Acquisition>
 WHERE
 {
      ?person foaf:name ?person_name .
      ?person  cf-coc:preserve ?media_name .
 }
```

As shown from above examples, query of RDF triples necessitate from juries the awareness of semantic parts and technical skills to write SPARQL code. In this case, the consumption of RDF data using SPARQL query language is not appropriate for juries to be one of their consumption patterns. Juries are specialized in law and legal procedures, not in the field of information technology. Thus, the need of a module that can reason over the RDF triples is required to be implemented. This will avoid that the juries need to be aware about this technical knowledge and proficient the SPARQL query code. This module will be based on different semantic rules of RDFS and the primitives of OWL.

SPARQL query language can not only query explicitly the RDF triples, but it is also able to infer triples that are not physically stored. This advantage resides on SPARQL when the latter has a rule base that can be used to infer implicit and hidden information.

In our context, LD is lightweight ontology using RDFS and some primitives of OWL. RDFS has some inference rules and reasoning for its constructors (see Table IV).

For example, the screen below shows a reasoning on owl:FunctionalProperty and owl:InverseFunctionalProperty of proprietary terms ‘preserve’ and ‘SN’, respectively.

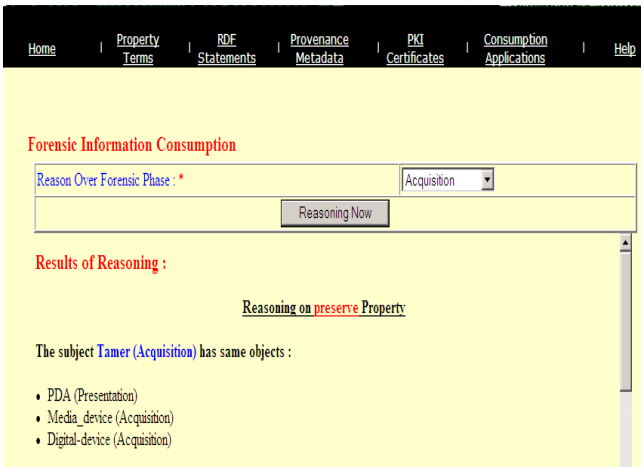


Figure 29. Reasoning on preserve Property

As shown in Figure 29, the juries select the ontology that he wants to reason. In this figure, they selected the acquisition phase (i.e., ontology), which contains the

property term called ‘preserve’ (i.e., the property of this term is tagged to be FunctionalProperty).

By referring to Table VI, we notice that if a property p is tagged to be FunctionalProperty, then all objects containing the predicate ‘preserve’ and having same subjects (i.e., Tamer), are equivalent to each other (i.e., PDA, Media_device, and Digital_device are equivalents).

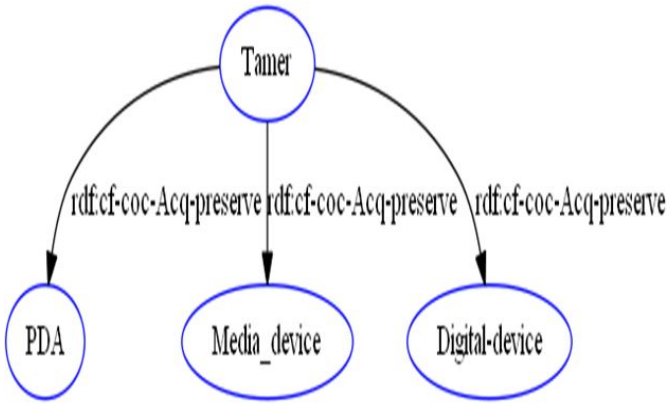


Figure 30. RDF Triples of preserve Property

Resource URL : https://127.0.0.1/Vocab/Acquisition#preserve

Term Type :	Property , Functional Property
Sub Property of :	made
Domain :	First_Responder
Range :	Digital_media
Inverse of :	preservedby

Below are the instances of class First_Responder	Predicate	Below are the instances of class Digital_media
Tamer (Acquisition Phase)	cf-coc-Acq-preserve	PDA (Presentation Phase)
Hamdi (Investigation Phase)	cf-coc-Acq-preserve	Computer (Acquisition Phase)
Tamer (Acquisition Phase)	cf-coc-Acq-preserve	Media_device (Acquisition Phase)
Hamdi (Investigation Phase)	cf-coc-Acq-preserve	Laptop (Acquisition Phase)
Tamer (Acquisition Phase)	cf-coc-Acq-preserve	Digital-device (Acquisition Phase)

Figure 31. Deferenceability of preserve Property

Again, by referring to Table VI, we notice that if a property ‘p’ is tagged to be InverseFunctionalProperty, then all subjects containing the predicate SN and having same objects (i.e., T1-236-185F), are equivalent to each other (i.e., Iphone and PDA, see Figures 32 and 33).

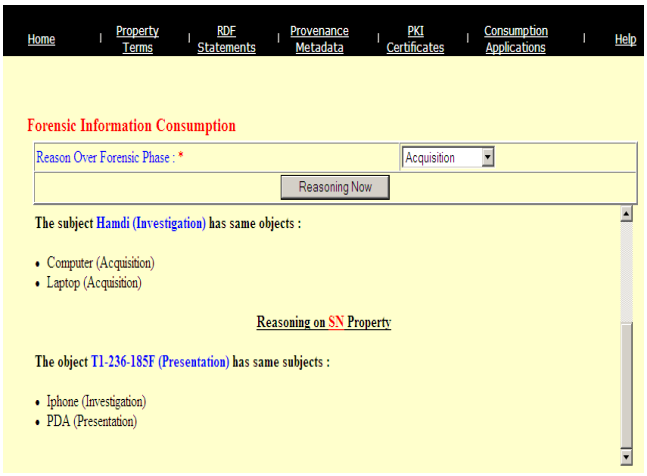


Figure 32. Reasoning on SN Property

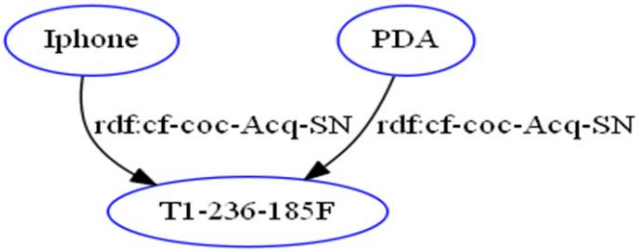


Figure 33. RDF Triples of SN Property

VI. CONCLUSION

This paper depicted a novel framework that will be used by role players to represent tangible *CoC* resulted from their cyber investigation. Role players use this framework to represent and publish forensic resources in order to be consumed by juries in the court of law.

This work explained in details all layers of the framework that are based on the technology stack of the LD. This technology stack (RDF, HTTP, and URI) is used to represent and publish different resources on the web in a structured way.

The role players start their representation process by defining new proprietary/custom terms describing the forensic information of their tangible *CoC*. This task is performed using lightweight ontology through the RDFS constructors and some primitive from OWL. Role players may also accompany different provenance metadata imported from the vocabularies of the semantic web to describe the origin of forensic information and strength the trustworthiness with the juries of court. All represented resources are then published in RDF format upon URI resolution, in order to be shared in a small scale between the role players and juries through the public key infrastructure approach. The latter opens the door to a new era of research representing the counter part of the linked open data, called

the linked closed data, which share all the advantages of the LOD, but with consumption restriction.

This work elaborated and explained the framework through the preservation task imported from the acquisition phase of the Kruse model. In future work, the framework will obey empirical experimentations through different scenarios imported from different forensic phases. This will be accomplished by defining all tasks for each phase of a forensic process. Different scenarios can be provided from different forensic models. In addition, supplementary layers may be added to this framework to facilitate communication between users (i.e., role players and users) and the framework. For example, the addition of an intelligent layer that transforms data between end-user and data store, and intelligent tutor layer that can guide the role players to use different well defined semantic vocabularies helps to define proprietary terms and their constraints, and learns role players the way of publishing data using such vocabularies. (i.e., in case they do not have enough technical knowledge about the LD).

REFERENCES

[1] T. F. Gayed, H. Lounis, and M. Bari, "Linked closed data using PKI: a case study on publishing and consuming data in a forensic process," The Sixth International Conference on Advanced Cognitive Technologies and Applications (IARIA 2014), Venice, Italy, pp. 77-86, ISBN: 978-1-61208-340-7.

[2] E. Kenneally, "Gatekeeping out of the box: open source software as a mechanism to assess reliability for digital evidence," Virginia Journal of Law and Technology, vol. 6, no. 13, issue 3, Fall 2001.

[3] T F. Gayed, H. Lounis, and M. Bari, "Representing and (im)proving the chain of custody using the semantic web," The Fourth International Conference on Advanced Cognitive Technologies and Applications (IARIA 2012), Nice, France, pp. 19-23, ISBN: 978-1-61208-218-9.

[4] M. W. Andrew, "Defining a process model for forensic analysis of digital devices and storage media," Proceedings of the 2nd International Workshop on Systematic Approaches to Digital Forensic Engineering (SADFE 2007), pp. 16-30, ISBN: 0-7695-2802-2, Seattle, WA.

[5] M. Köhn, J. H. P. Eloff, and Ms. Olivier, "UML modeling of digital forensic process models (DFPMs)," Proceedings of the International Security South Africa (ISSA 2008), Innovative Minds Conference, Johannesburg, South Africa, pp. 1-13, Jul. 2008.

[6] Electronic Crime Scene Investigation: A Guide for First Responders, Second Edition, 78 Pages, by Independent Publishing Platform; 2 Edition (July 9, 2012), ISBN-10: 147827683, ISBN-13: 878-1478276845.

[7] E. Casey, "Digital Evidence and Computer Crime - Forensic Science Computers and the Internet," 3rd Edition, Academic Press 2011, pp. 1-807, ISBN: 978-0-12-374268-1.

[8] S. O. Ciardhuain, "An extended model of cyber crime investigations," International Journal of Digital Evidence, vol. 3, issue 1, 2004.

[9] Y. Yusoff, R. Ismail, and Z. Hassan, "Common phases of computer forensics investigation models," International Journal of computer science and information technology (IJCSIT 2011), vol. 3, no. 3, pp. 17- 31.

[10] T. F. Gayed, H. Lounis, and M. Bari, "Computer forensics: toward the construction of electronic chain of custody on the semantic web," Proceedings of the 24th International Conference on Software Engineering & Knowledge Engineering (SEKE 2012), pp. 406-411.

[11] T. F. Gayed, H. Lounis, and M. Bari, "Cyber forensics: representing and managing tangible chain of custody using the linked data principles," The international conference on Advanced Cognitive technologies and Application (IARIA 2013), pp. 87-96, ISSN: 2308-4197, ISBN: 978-1-61208-273-8, Valencia, Spain.

[12] T. F. Gayed, H. Lounis, and M. Bari, "Representing chains of custody along a forensic process: a case study on Kruse model," Proceedings of the 25th International Conference on Software Engineering & Knowledge Engineering (SEKE'2013), pp. 674-680, ISBN 1-891706-33-0 ISSN: 2325-9000.

[13] G. C. Kessler, "Judges' Awareness, Understanding and Application of Digital Evidence," PhD Thesis in Computer Technology in Education, Graduate school of computer and information sciences, Nova Southeastern University, 2010.

[14] RDF Model and Syntax Specification, W3C recommendation, 22 Feb 1999, www.w3.org/TR/REC-rdf-syntax-19990222/1999 [retrieved Oct. 2014].

[15] D. Brickley and R. V. Guha. "RDF Schema," W3C Recommendation http://www.w3.org/TR/rdf-schema/ [retrieved Nov. 2014].

[16] OWL: web ontology language overview, W3C Recommendation, http://www.w3.org/TR/owl-features/ 2004 [retrieved Nov. 2014].

[17] O. Hartig, "Provenance information in the web of data," In proceedings of the linked data on the web (LDOW 2009), Workshop at the World Wide Web Conference (WWW), Madrid, Spain.

[18] Dublin Core Metadata Initiative: http://dublincore.org/documents/dcmi-terms/ [retrieved Nov. 2014].

[19] FOAF Vocabulary Specification 0.99: http://xmlns.com/foaf/spec/ [retrieved Nov. 2014].

[20] P. P. da Silva, D. L. McGuinness, and R. Fikes, "A proof markup language for semantic web services," Information Systems, pp. 381–395, vol. 31, issue 4, Jun. 2006.

[21] T. Berners-Lee, R. Fielding, and L. Masinter, "RFC 2396 – Uniform Resource identifiers (URI): Generic Syntax," http://www.isi.edu/in-notes/rfc2396, Aug 1998 [retrieved Feb. 2013].

[22] R. Fielding, "Hypertext transfer protocol," – http/1.1 request for comments:http://www.w3.or/Protocols/rfc2616/rfc2616.html, 1999 [retrieved Mar. 2013].

[23] T. Omitola et al., "Tracing the provenance of linked data using voiD," The International Conference on Web Intelligence, Mining and Semantics (WIMS 2011), vol. 17, no. 17, ISBN: 978-1-4503-0148-0.

[24] L. T. Berners-Lee, "Design issues: linked data," from http://www.w3.org/DesignIssues/LinkedData.html [retrieved Feb. 2013].

[25] M. Cobden, J. Black, N. Gibbins, L. Carr, and N. R. Shadbolt, "A research agenda for linked closed dataset," In Proceeding of the 2nd International Workshop on Consuming Linked Data (COLD 2011), vol. 782 [vision paper].

[26] Linking Open Data, W3C Semantic Web Education and Outreach (SWEO) Community Project, http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData [retrieved: Oct. 2014].

[27] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data— The story so far," International Journal on Semantic Web and Information Systems (IJ-SWIS 2009), vol. 5, no. 3, pp. 1-22.

[28] T. F. Gayed, H. Lounis, and M. Bari, "Creating proprietary terms using lightweight ontology: a case study on acquisition phase in a cyber forensic process," Proceedings of the 26th International Conference on Software Engineering & Knowledge Engineering (SEKE 2014), Vancouver, Canada, pp. 76-81.

[29] T. Berners-Lee, J. Hendler, and Ora Lassilia, "The semantic web," Scientific American, vol. 5, May 2001, pp. 34–44.

[30] RDF/XML Specifications: W3C Recommendation 10 February 2004, http://www.w3c.org/TR/REC-rdf-syntax [retrieved Oct. 2014].

[31] Turtle – Terse RDF Triple Language: W3C Team Submission 14 January 2008, http://www.w3.org/TeamSubmission/turtle/ [retrieved Nov. 2014].

[32] Resource Description Framework Anchor (RDFa) in Extensible Hyper Text Markup Language (XHTML): Syntax and Processing,

W3C Recommendation 14 Oct. 2008, http://www.w3.org/TR/rdfa-syntax/ [retrieved Nov 2014].

[33] RDF 1.1 Test Cases: http://www.w3.org/TR/2014/NOTE-rdf11-testcases-20140225/ [retrieved Nov. 2014].

[34] Notation3 (N3): A Readable RDF Syntax: W3C Team Submission 14 January 2008, http://www.w3.org/TeamSubmission/n3/ [retrieved Sep. 2014]

[35] L. M. Campbell and S. MacNeill, "The semantic web, linked and open data, a briefing paper," Centre for Educational Technology, Interoperability and Standards (JISC CETIS 2010), SN: 2010:B03.

[36] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao, "Describing Linked Datasets - On the Design and Usage of VoiD, 'the Vocabulary Of Interlinked Datasets'," WWW 2009 Workshop: Linked Data on the Web (LDOW 2009), Madrid, Spain.

[37] P. Hitzler and F. V. Harmelen, "A reasonable semantic web," vol. 1(1), 2010, pp. 39-44, URL: http://corescholar.libraries.wright.edu/cse/25 [retrieved Sep. 2014].

[38] G. Antoniou and F. V. Harmelen, "Web ontology language: OWL," Handbook on Ontologies 2004, pp. 67-92.

[39] S. Vanstone, Oorschot, and A. Menezes, "Handbook of Applied Cryptography," CRC Press, 1997, ISBN: 0-8493-8523-7.

[40] C. LT. Brown, "Computer Evidence: Collection & Preservation," 1ST Edition, ISBN: 0-619-13120-9, 2006.

[41] T. Health and C. Bizer, "Linked Data: Evolving the Web into a Global Data Space," http://linkeddatabook.com/editions/1.0/ [retrieved: Nov. 2014].

[42] D. Berrueta and J. Phipps, "Best Practice Recipes for Publishing RDF vocabularies," - w3c note. http://www.w3.org/TR/swbp-vocab-pub/, 2008 [retrieved Nov. 2014].

[43] N. Mendelsohn, "The Self-Describing Web," http://www.w3.org/2001/tag/doc/selfDescribingDocuments.html, 2009 [retrieved Sep. 2014].

[44] A. Brinson, A. Robinson, and M. Rogers, "A Cyber Forensics Ontology: Creating a New Approach to Studying Cyber Forensics," The International Journal of Digital Forensics & Incident Response (DFRWS 2006). vol. 3, pp. 37-43.

[45] Simple Knowledge Organization System: http://www.w3.org/2004/02/skos/ [retrieved Oct. 2014].

[46] W. Kruse II and J. Heiser, "Computer Forensics: Incident Response Essentials," Addison Wesley, 2002, ISBN-13: 978-0201707199, ISBN-10: 0201707195, 1st Edition.

[47] OWL Web Ontology Language Guide: http://www.w3.org/TR/owl-guide/ [retrieved: Nov. 2014]

[48] G. Palmer, "A road map for digital forensic research," Technical Report from the First Digital Forensic Research Workshop (DFRWS), Utica, New York, 2001, DTR – T001-01 Final.

[49] B. D. Carrier, "A Hypothesis-based approach to digital forensic investigations," PhD thesis, Center for Education and Research in Information Assurance and Security, Purdue University, West Lafayette, IN 47907-2086.

[50] B. Carrier, "Defining digital forensic examination and analysis tool using abstraction layers," International Journal of Digital Evidence IJDE 2003, vol. 1, issue 4, pp. 1-12.

[51] E. Politcnica, "Improving chain of custody in forensic investigation of electronic digital systems," International Journal of Computer and Networks Security (IJCNS 2011), vol. 11, no. 1, pp. 1-9.

[52] J. Cosic and M. Baca, "A framework to (im)prove chain of custody in digital investigation process," Proceedings of the 21st Central European Conference on Information and Intelligent Systems (CECIIS 2010), pp. 435-438, Varaždin, Croatia.

[53] J. Cosic and M. Baca, "(Im)proving chain of custody and digital evidence integrity with timestamp," 33rd Proceedings of the International Convention on Information and Communication Technology Electronics and Microelectronics, ICT Convention (MIPRO 2010), Opatija, pp. 1226-1230, ISBN: 978-1-4244-7763-0.

[54] R. Jueneman and R. Lapedis, "Solving the digital chain of custody problem," Trusted Mobility Solutions, SPYRUS 2010, Document number 412-000001-02 [white paper].

[55] A. Bogen and D. Dampier, "Knowledge discovery and experience modeling in computer forensics media analysis," In International Symposium on Information and Communication Technologies (ISICT 2004), pp. 140-145, ISBN: 1-59593-170-8.

[56] B. Schatz, PhD thesis, "Digital Evidence: Representation and Assurance," Faculty of Information technology, Queensland University of Technology, Oct. 2007.

[57] B. Schatz, G. Mohay, and A. Clark, "Rich event representation for computer forensics," Proceedings of the 17TH International Conference in Knowledge Based and Intelligent Information and Engineering Systems (KES 2013), vol. 22, pp. 1266-1275, ISBN: 0-9596291-9-1.

[58] B. Schatz, G. Mohay, and A. Clark, "Generalising event forensics across multiple domains," Proceedings of the 2004 Australian Computer Network and Information Forensics Conference (ACNIFC 2004), pp. 136-144, Perth, Australia.

[59] S. Al-Fedaghi and B. Al-Babtain, "Modeling the forensic process," International Journal of Security and its Application, Vol. 6, no. 4, Oct. 2012, pp. 79-107.

[60] Common Digital Evidence Format (CDESF), Available from: http://www.dfrws.org/CDESF/index.html [retrieved Nov. 2014].

[61] S. L. Garfinkel, D. J. Malan, K-A. Dubec, C.C. Stevens, and C. Pham, "Disk imaging with the advanced forensics format, library and tools," Advances in Digital Forensics, 2nd Annual IFIP WG 11.9, In Proceeding of International Conference on Digital Forensics 2006, Orlando, Florida.

[62] P. Turner, "Unification of Digital Evidence from Disparate Sources (Digital Evidence Bags)," In 5th Digital Forensic Research Workshop (DFRW 2004), vol. 2, issue 3, pp. 223-225, New Orleans.

[63] M. Cohen, M. Garfinkel, and B. Schatz, "Extending the advanced forensic format (AFF) to accommodate multiple data sources, logical evidence, arbitrary information and forensic workflow," Digital Investigation: The International Journal of Digital Forensics and Incident (IJDFI 2009), vol. 6, pp. 57-68, ISSN: 1742-2876.

[64] Simple Knowledge Organization System RDF Schema (SKOS) Vocabulary: http://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html [retrieved Nov. 2014].

[65] Open Provenance Model Vocabulary Specification: http://purl.org/net/opmv/ns [retrieved Oct. 2014].

[66] O. Hartig and J. Zhao, "Publishing and consuming provenance metadata on the web of linked data," International Provenance and Annotation Workshop (IPAW 2010), LNCS 6378, vol. 6378, Berlin, pp. 78-90, ISBN 978-3-642-17819-1.

[67] How to Publish Linked Data on the Web: http://www4.wiwiss.fu-berlin.de/bizer/pub/linkeddatatutorial/ [retrieved Oct. 2014].

[68] L. Moreau et al., "The open provenance model core specification," (v1.1), Future Generation Computer. Systems., vol. 27, issue 6, 2011, pp. 743-756.

[69] J. J. Carroll, C. Bizer, P. Hayes, and P. Stickler, "Named graphs, provenance and trust," In Proceeding of WWW '05 of the 14th international conference on World Wide Web (WWW), USA, ACM Press, pp. 613-622, ISBN: 1-59593-046-9.

[70] R. Isele, A. Harth, J. Umbrich, and C. Bizer, "Ldspider: An open-source crawling framework for the web of linked data," In 9th International Semantic Web Conference (ISWC 2010) & Demonstrations Trash: Collected Abstracts vol. 658.

[71] Describing Linked Datasets with the VoID: http://www.w3.org/TR/void/ [retrieved Mar. 2014].

[72] E. Rajabi, M. Kahani, and M. Angel Silicia, "Trustworthiness of Linked Data Using PKI," World Wide Web Conference (www2012) Lyon, France, 2012.

[73] J. Davies, "Implementing SSL/TLS Using Cryptography and PKI," Indianapolis, Indiana: Wiley Publishing Inc, ISBN: 978 -0-470-92041-1, 2011.

[74] D. Richard, V. C. Hu, W. Timothy, and S. Chang, "Introduction to public key technology and the federal PKI infrastructure," Technical Report, SP 800-32, National Institute of Standards and Technology (NIST), U.S. Government publication, 2013 Edition.

[75] M. Blaze, J. Feigenbaum, and A. Keromytis, "KeyNote: Trust management in the public key infrastructure," 6th International Workshop Cambridge, UK, vol. 1550, pp. 59-63, ISBN: 978-3-540-49135-4 January 1999 [White paper].

[76] E. Barker et al., "Recommendation for Key Management Part 3: Application-Specific Key Management Guidance," NIST Special Publication 800-57, 2013 Edition.

[77] Public Key Infrastructure, Entrust: www.entrust.com/what-is-pki/ [retrieved: Feb. 2014].

[78] R. Perlman, "An overview of PKI trust models, In IEEE network," vol. 13, issue 6, pp. 38-43,1999, ISSN: 0890-8044.

[79] Extended Validation SSL Certificate: The Next Generation High Assurance SSL Certificate, http://www.evsslcertificate.com/ssl/description-ssl.html [retrieved: Sep. 2014].

[80] Internet X.509 Public Key Infrastructure Certificate Management Protocols: https://tools.ietf.org/html/rfc2510 [retrieved: Mar. 2014].

[81] Official Site of OpenSSL Project, http://www.openssl.org/ [retrieved: Dec. 2013].

[82] Cyber Forensics-Chain of Custody Server Host, Domain owned by Tamer Gayed, www.cyberforensics-coc.com [retrieved: Oct. 2013].

[83] D. Quan and D. R. Karger, "How to make a semantic web browser," WWW'04 Proceedings of the 13th International Conference on World Wide Web, pp. 255-265, New York, USA, ISBN: 1-58113-844-X.

[84] Semantic Web Search Engine (SWSE): http://www.swse.org/ [retrieved Oct. 2014].

[85] Semantic Web Search (Swoogle): http://swoogle.umbc.edu/ [retrieved Sep. 2014].

[86] L. Davis, H. Shrobe, and P. Szolovits, "What is a knowledge representation?," AI Magazine, 1993, vol. 14(1), pp. 17-32.

[87] J. Zhao, C. Bizer, Y. Gil, P. Missier, and S. Sahoo, "Provenance requirements for the next version of RDF," A position paper based on the work of the W3C Provenance Incubator Group, Standford, CA, June 2010.

[88] B. Glimm, A. Hogan, M. Krötzsch, and A. Polleres, "OWL: yet to arrive on the web of data?," In Proceeding of Linked Data on the Web Workshop (LDOW 2012), vol. 937, Lyon, France.

[89] A. Polleres, A. Hogan, R. Delbru, and J. Umbrich "RDFS & OWL Reasoning for Linked Data," Semantic Technologies for Intelligent Data Access, Lecture Notes in Computer Science, vol. 8067, pp. 91-149. ISBN: 978-3-642-39783-7.

[90] Easy RDF: http://www.easyrdf.org/ [retrieved: Nov. 2014].

[91] Graphiz – Graph Visualization Software: http://www.graphviz.org/Documentation.php [retrieved: Feb. 2013].

[92] Internet Information Services, Microsoft, http://www.iis.net/ [retrieved: Mar. 2014].

[93] Official Site of OpenSSL Project, http://www.openssl.org/ [retrieved: Dec. 2013].

[94] T. Heath, "How will we interact with the web of data?," IEEE Internet Computing 2008, vol. 12, issue 5, pp. 88–91, ISSN: 1089-7801.

[95] T. Berners-Lee et al., "Tabulator: Exploring and analyzing linked data on the semantic web," In Proceedings of the 3rd International Semantic Web User Interaction Workshop (SWUI 2006), Athens, Georgia.

[96] O. Hartig, C. Bizer, and J. Christoph Freytag, "Executing SPARQL queries over the web of linked data," In Proceedings of the 8th International Semantic Web Conference (ISWC 2009), vol. 5823, pp. 293–309, ISBN: 978-3-642-04930-9.

[97] G. Cheng and Y. Qu, "Searching linked objects with falcons: Approach, implementation and evaluation," International Journal on Semantic Web and Information Systems (IJSWIS 2009), pp. 49–70.

[98] http://inference-web.org/2007/primer/ [retrieved: Jan. 2013].

[99] Server Certificate Installation Instructions, Microsoft Developer Network: http://msdn.microsoft.com/en-us/library/ms751408.aspx [retrieved: Feb. 2014].

[100] IIS Management Microsoft Management Console (MMC), Microsoft, http://support.microsoft.com/kb/892987 [retrieved: Feb. 2014].

Modelling Spatial Understanding: Using Knowledge Representation to Enable Spatial Awareness and Symbol Grounding in a Robotics Platform

Martin Lochner, Charlotte Sennersten, Ahsan Morshed, and Craig Lindley
CSIRO Computational Informatics (CCI) Autonomous Systems (AS)
Commonwealth Scientific and Industrial Research Organization (CSIRO)
Hobart, Tasmania, Australia
Contact: martin.lochner@csiro.au, charlotte.sennersten@csiro.au,
ahsan.morshed@csiro.au, craig.lindley@csiro.au

***Abstract*—Robotics in the 21st century will progress from scripted interactions with the physical world, where human programming input is the bottleneck in the robot's ability to sense, think and act, to a point where the robotic system is able to autonomously generate adaptive representations of its surroundings, and further, to implement decisions regarding this environment. A key factor in this development will be the ability of the robotic platform to understand its physical space. In this paper, we describe a rationale and framework for developing spatial understanding in a robotics platform, using knowledge representation in the form of a hybrid spatial-ontological model of the physical world. Further, we describe the proposed CogOnto (cognitive ontology) model, which enables symbol grounding for a cognitive computing system, using sensor data gathered from diverse and heterogeneous sources, associated with humanly crafted symbolic descriptors. While such a system may be implemented with classical ontologies, we discuss the advantages of non-hierarchical modes of knowledge representation, including a conceptual link between information processing ontologies and contemporary cognitive models.**

Keywords-Human Robot Interaction; Artificial Intelligence; Autonomous Navigation; Knowledge Representation; Symbol Grounding; Spatial Ontology.

I. INTRODUCTION

The process of transitioning away from hard-coded robotics applications, which carry out highly pre-determined actions such as the traditional manufacturing robot, is already well underway. This paper follows our previous work [1] in which we describe a methodology for using ontological data representation to encode 3D spatial information in robotics applications. With notions such as cloud robotics [2] entering the *zeitgeist*, and highly publicized events such as the Defense Advanced Research Projects Agency (DARPA) Robotics Challenge (Dec. 19-21, 2013, Miami FL) bringing public attention to these advances, it is foreseeable that robots will be entering the mainstream realm of human activity – more than in fringe applications (robotic vacuum cleaner; children's toys), but in key areas such as caring for the aged [3], operating vehicles [4], disaster management [5], and undertaking autonomous scientific investigation [6].

The hurdles that must be overcome in reaching these goals, however, are neither few nor small. This can be plainly seen, for example in the aforementioned 2013 Robotics Challenge, in which simple spatial tasks that are routine for a human being (open a door, climb a ladder) are still critically difficult for even the most advanced and highly funded robotics projects. While the state-of-the-art is impressive, it is evident that physical robotics hardware is far in advance of the control systems that are in place to guide the robot. The challenge is, thus, to develop systems whereby a robot can perceive a physical space and understand its position in that space, the components that exist within the space, and how it can or *should* interact with these components in order to achieve implicit or explicit goals. This is furthermore impacted by the requirement that robotic systems be able to operate in outdoor environments where distributed connections may not be available; however, describing the development of long-range data networks for robotic communication is beyond the scope of this paper.

While there are a number of ways that the problem of providing a robot with a spatial understanding can be approached (e.g., neuro-fuzzy reasoning [7], dynamic spatial relations via natural language [8]) it is our proposition that leveraging the current advancements in knowledge representation via ontologies [9][10], in combination with an understanding of human spatial-cognitive processing [11][12], and enabled by real-time scene modeling [13] will provide a powerful and accessible methodology for enabling spatial understanding and interaction in a mobile robotics platform. As argued by Sennersten et al. [14], the advantage of using cloud-based repositories of perceptual data annotated with ontology and metadata information is to take advantage of humanly-tagged examples of sense data (e.g., images) to overcome the symbol grounding problem. Symbol grounding refers to the need for symbolic structures to have valid associations with the things in the world that they refer to. Achieving symbol grounding is an ongoing challenge for robotics and other intelligent systems [15]. Using cloud-based annotations attached to sensory exemplars takes advantage of the human ability to ground symbols, obviating the need

for robots to achieve this independently of human symbolic expressions.

This paper provides a conceptual overview of how spatial understanding can be developed in a robotics platform. We discuss traditional knowledge representation (classical information processing ontologies), describe the development and use of "cognitive" ontologies, and how this may be transitioned into the development of a physical-spatial ontology, including a possible system of comprehension for spatial position. Finally, we discuss the notion that truly non-hierarchical systems such as complex chemical structure, and such as the human cortex, may require the development of systems of knowledge representation that transcend the structural limits of today's systems.

II. STATE OF THE ART: KNOWLEDGE REPRESENTATION

The development of specific nomological hierarchies for concept representation is currently taking place across many fields of academic endeavor (e.g., genetics, medicine, neuroscience, biology, chemistry, physics). Under the guise of the philosophical concept of an *Ontology*, such applications seek to outline the knowledge, which exists within a domain at three levels of representation: Classes, Properties, and Relationships. These nomological hierarchies provide a way of describing the precise relationship that terms in a given domain have to one another. As an information processing construct, the definition of an ontology is refined as an "explicit formal specification of the terms in the domain and relations among them", or more concisely, "a specification of a conceptualization" [16].

A system that operates with such knowledge representation within its core functionality may be considered to be 'knowledge-based'. A knowledge-based system is a computer program that stores knowledge about a given domain (also known as an "expert system", when the knowledge is considered to be from a highly specialized domain). However, an ontology does not intrinsically represent the kinds of truth-functional mappings or procedures captured by rules in more complete knowledge bases. Hence, an ontology provides classifications and the ability to infer associations via subclass/superclass relationships. More complex forms of reasoning required for most forms of useful cognitive task performance require task-oriented rules. As such, the domain knowledge in a knowledge base includes ontology representations, while most task-oriented reasoning is achieved by the use of rules that refer to ontological constructs in the form of domains within rule tuples.

The system attempts to mimic the reasoning of a human specialist by conducting reasoning across rules and in reference to a database of atomic facts. Matching sense data against metadata/ontology-annotated sense data on the web can provide a method of automatically mapping a current sensed situation to the annotations of past situations stored in the cloud. This allows the system to retrieve representations of the situation in an atomic form, as statements formulated using the symbolic forms of annotations, which are retrieved by matching against associated sense data. Ontologies hold the potential, therefore, to provide the constructs for symbolic atomic fact expressions that rule-sets can then process for automated cognitive task performance.

A. Cognitive Ontologies

An increasing number of ontologies are available on-line that can potentially support this symbolic structure generation process. Knowledge representation via ontological structure has been applied to the field of cognitive science, both in relation to terminology used within the domain (e.g., DOLCE - Descriptive Ontology for Linguistic and Cognitive Engineering [17][18]) and for concepts relevant to empirical testing paradigms (e.g., CogPo [19]). Indeed, several cognitive ontologies have been developed in the recent years, including DOLCE, WordNet [20], CYC [21], and CogPo.

WordNet is an online lexical knowledgebase system, whose design is inspired by current psycholinguistic theories of human lexical memory, where each cognitive artifact can be semantically classified into English nouns, verbs, and adjectives, with different meanings and relationships in real-world scenarios. DOLCE is developed by Nicola Guarino and his associates at the Laboratory for Applied Ontology (LOA) [22]. It captures the ontological categories underlying natural language and human common sense. DOLCE, however, does not commit to a particularly abstract level of concepts that relate to the world (like imaginary thoughts); rather, the categories it introduces are thought of as cognitive artifacts, which are ultimately dependent on human perception, cultural imprints and social conventions.

The Cyc project goal is to build a larger common-sense background knowledgebase, which is intended to support unforseen future knowledge representation and reasoning tasks. The Cyc knowledgebase contains 2.2 million assertions (fact and rules) describing more than 250,000 terms, including nearly 15,000 predicates.

Finally, the Cognitive Paradigm Ontology (CogPo) is developed based on two well-known databases, namely, the Functional Imaging Biomedical Informatics Research Network (FBIRN) Human Imaging Data base [23] and the BrainMap database [24]. The CogPo Ontology has categorized each paradigm in terms of **(**1) the stimulus presented to the subjects, (2) the requested instructions, and

(3) the returned response. All paradigms are essentially comprised of these three orthogonal components, and formalizing an ontology around them is a clear and direct approach to describing paradigms. This well-formed standard ontology guides cognitive experiments in formalizing the cognitive knowledge.

While these ontologies are of great value to the community of researchers, and while the knowledge-based mapping of concepts within particular domains may enable robotic systems to rapidly access the linguistic identity of physical objects and their relations within the domain, they do not provide a means whereby the robot may become spatially aware. To achieve this goal, we will need to provide the robot with the ability to identify the spatial characteristics particular to an identified object, and the physical relations between these objects and the surrounding environment. A robot requires an internal representation of 3D space. It could access two dimensional images on the web, by content-matching those images with contents of its own visual system. This would aid the robot by enabling real-time identification of unfamiliar objects, including spatial parameters that may not be immediately visible to on-board sensors. The matching process, and especially the ongoing 3D interpretation of the images, could be greatly aided if the ontology/metadata associated with images includes representation of the 3D context of image capture. The "ontological" schema of knowledge representation for images may provide this means if it is extended to include 3D spatial annotations.

III. REPRESENTING RELATIONSHIPS IN THREE DIMENSIONS: SPATIAL ONTOLOGIES

We propose here that this same methodology for specifying semantic relationships between concepts (the ontological structure of knowledge representation, i.e., Classes, Properties, and Relationships) may also be useful in specifying spatial relationships between physical objects. While a traditional ontology will hierarchically represent a concept and its relation to other concepts in a domain, a spatial ontology (e.g., Fig. 1) will represent an object, (class), its spatial properties including a detailed 3D representation in a language such as the X3D XML-based file format, and its positional relation (x,y,z) to other objects existing within the scene by using the datatype properties.

An entity (the "*individual*") in a prototypical ontology is comparable to an entity in a spatial ontology, being an object in the physical world. *Class* indicates the category, into which the individual falls, for example "person", or "boat". *Attributes* traditionally describe the *individual* – features, properties, or characteristics of the object: a person has arms; a boat has a hull. In a spatial ontology this information will be appended with configural information regarding the object, for example the parent-child node relationship of a human body, including torso, appendages, etc. The *relation* between individuals is where the power of the traditional ontology arises, by specifying the precise ways, in which different individuals relate to one another (e.g., "a catamaran is a subclass of boat"). Once again, in a spatial ontology the *relation* will be a precise indicator (a reference, or an 'object index') of the relative positionality of items in the physical space, as described in the following section. By thus, leveraging the existing functionality of ontological representation, augmented with relevant and necessary spatial referencing information, we may develop a knowledge-based system that enables a level of spatial awareness in a robotic platform.

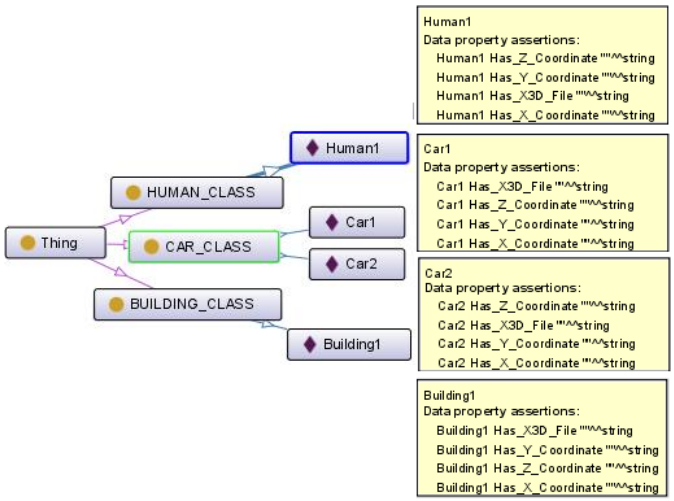


Figure 1. Example of a simple spatial ontology
(Note that the relations between objects are represented via "Data Properties" here.)

A. A system of comprehension for spatial position

Following the above discussion about relationships in 3D space, we look into how coordinate systems can be synchronized for objects whose positions and local configurations are non-static. The physical scale requirement that a robot needs to have can be measured by the accuracy the robot needs to operate in via its navigation system. An autonomous robot must be able to determine its position in order to be able to navigate and interact with its environment correctly (e.g., Dixon and Henlich, 1997 [25]). When the *Class* of "robot" navigates from A to B it is a basic motion, which is similar to the movement of an in-game character via a default keyboard set-up where the key "W" moves the character forward, turning left using key "A", turning right using key "D" and go backwards using key "Z". The 3D digital world uses the X, Y, Z coordinate system called the Cartesian Coordinate Method (CCM) and is expressed in meters (m). To measure distance between two spherical points; X^1, Y^1, Z^1 and X^2, Y^2, Z^2 we take the

Euclidean distance using a Cartesian version of Pythagoras' Theorem (1). The distance is the sum of their individual point differences in square.

$$\sqrt{((x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2)} \qquad (1)$$

To determine a position in the physical world and navigate the robot in map-referenced terms to a desired destination point from A to B, Dixon and Henlich use what they call 1) *Global Navigation*. The positioning accuracy with a standard consumer Geographical Positioning System (GPS) is accurate within a range of 8 feet, which is approximately 284 centimeters. This does not give high fidelity position accuracy. As such, when the robot has to operate in a typical indoor manufacturing environment, it needs detailed position support in order to create 3D reference points within the space. What Dixon and Henlich call 2) *Local Navigation* is to determine one's own position relative to the objects (stationary or moving) in the environment, and to interact with them correctly. If we think of Human Robot Interaction (HRI) and the robot arm and its gripper(s) (hand/s), the gripper(s) must via eye(s) be able to recognize the object it will manipulate and how it shall be manipulated. The spatial centre points for individual objects are of importance, as well as group of objects and the robot's own centre point in relation to actual manipulation centre point for gripper. From a spatial ontology point of view, the centre points have to be able to change dynamically depending on interaction purpose.

For example, the Puma robot arm series has three different arms with slightly different sophistication and these are Puma 200, Puma 500, and the Puma 700 Series. These robot arms execute 3) *Personal Navigation* [D&H], which makes the arm aware of the positioning of the various parts, its own positioning, and also in relation to each other and in handling objects. The Puma 200 Series has been used for absolute positioning accuracy for CT guided stereotactic brain surgery [26]. The Puma 200 robot has a relative accuracy of 0.05 mm. There are already 3D Spatial Vision Systems for robots out on the market, which are driven via several cameras. This creates a local world solution for 3D vision robot guidance, where the software first makes the user calibrate the cameras and the robot, and then loads standard Computer Aided Design (CAD) files of parts, which the system shall track.

IV. THE 3D WORLD

The ability to scan a real-world environment makes it possible to extract digital information about the physical world, and the way in which it functions. Three dimensional perception is a key technology for robotics applications where obstacle detection, mapping and localization are core capabilities for operating in unstructured environments. Laser scanning creates a surface point cloud of a 3D physical environment [34] making it possible to map any environment in a rather short time (the Leaning Tower of Pisa was scanned in 20 minutes). This technology can be used in a robotic intelligence system for Simultaneous Localization Mapping (SLAM) and higher level reasoning regarding location and position. However, object recognition and manipulation requires deriving 3D object information from the overall point cloud and building cognitive models with task reasoning for using object and scene data in real time.

Object extraction [35][36][37] makes it possible to know what a robot is looking at, supporting manipulation or collection actions. This can be achieved by an Environmental Scanning-Object Extraction (ES-OE) engine. For human-robot collaboration, a robot can be enabled to use deictic visual references from human gaze by integrating an eye tracker with the ES-OE engine.

A. Background

In a previous work [38], a 3D simulation engine was integrated with an eye tracker. The integrated system allows the human point of gaze on 3D objects within a 3D digital world projected onto a computer screen to be tracked automatically. This development made it possible to log gaze in various task-related environments in a simulated world. From a Human Factor's perspective, the simulation and human observation can be investigated, including collaborative actions performed by groups with various workloads, stressors and decisions. There have been several studies made using the technological framework with different stimuli [39][40][41], but no substantial theoretical framework has been developed in relation to this object-based approach *per se*. A bottleneck in relation to this visual approach has been that 2D image, film and visual stimuli have not met the requirements for incorporating a knowledge-based approach for dynamic 3D worlds, whether the real physical world or a digitized 3D world. The object approach needs to address how both modeled and real world objects can be perceived and manipulated [42] by a robot, allowing the system to sense, think and act in real time: the computer needs to understand how to define an object and how to ontologically and semantically make sense out of such an object in a dynamic spatial world.

1) 3D objects in a 3D world

In [38], a simulation engine integrated with an eye tracker took a gaze fixation (x and y screen coordinates) and ray casted/traced from that position onto the underlying 3D virtual object's collision box, a volume corresponding with the shape of a virtual object as recognized and processed by a physics engine that is also used to designate objects by interface devices, like a mouse. This made it possible to track gazed objects in real time every 17 ms (using a 60Hz eye tracker). The same principle can be used in a physical world context where an ES-OE engine could be integrated with eye tracking glasses to allow a computational system to know what object a person wearing the glasses is looking at.

2) Structuring a noisy world

The 3D world scenario, simulated or physically real, constitutes an event or scene. A scenario includes objects that are instances of their classes. A class could be something like a *CarClass*, *HumanClass*, *FlowerClass*, etc.

In a constrained world, we can name all objects beforehand so when they are logged we know what they are and what position (x, y, z, θ_1, θ_2, θ_3) they are in. In an unconstrained environment that is scanned and has extracted objects, we must also have a capability to know what the objects are and to be able to classify them. A cloud-based approach of the kind proposed in this paper presents a middle ground, being more open than a highly constrained environment, but still being limited to objects of types that are represented and labeled within the cloud.

V. INTELLIGENT ACTION IN A STRUCTURED WORLD

Knowledge by definition is "1. Facts, information, and skills acquired through experience or education; the theoretical or practical understanding of a subject and 2. Awareness or familiarity gained by experience of a fact or situation" [43]. To gain an understanding of how robots might learn and operate on knowledge, we have looked at several established models that can fit within an initial architecture that enhances these established models by the ingestion of information from the web. Our overall aim is to build a computational comprehension system for 3D object information, assisted by a hybrid computational ontology (i.e., combining several existing and new ontologies).

A. Existing Models

Extensive effort has been put into the task of understanding and attempting to re-create/simulate the processes, by which a human being thinks. Using the underlying assumption that intelligence is wholly "the simple accrual and tuning of many small units of knowledge" [44], production-based models of cognition have had success in displaying human-like performance on a number of tasks (e.g., visual search [45] and natural language processing [46]). While there are debates regarding the similarity of what humans actually do to what we have achieved using the above assumption [47], there is little doubt that such systems can produce intelligent-seeming behavior, which can facilitate the development of vitally useful control structures in the field of robotics and computational intelligence [46].

One of the most influential models of human cognition is the ACT-R, or "Adaptive Character of Thought – Rational" model [44], developed over many years by John Anderson, who was a student of the seminal Cognitive Scientist Alan Newell (1927-1992). Anderson's model is a hybrid symbolic/sub-symbolic system that incorporates various "modules" that are deemed necessary for rational behavior, and are thought to have biological correlates. These include the modules *Declarative* (manages creation, storage and activation of memory "chunks"), *Procedural* (stores and executes productions based on expected utility), *Intentional/Imaginal* (goal formulation for directed behavior), and *Visual (2D)/Audio* (theoretically plausible implementation of visual and auditory perception), see Fig. 2. An internal pattern-matching function searches for a production that matches the current state of the buffers.

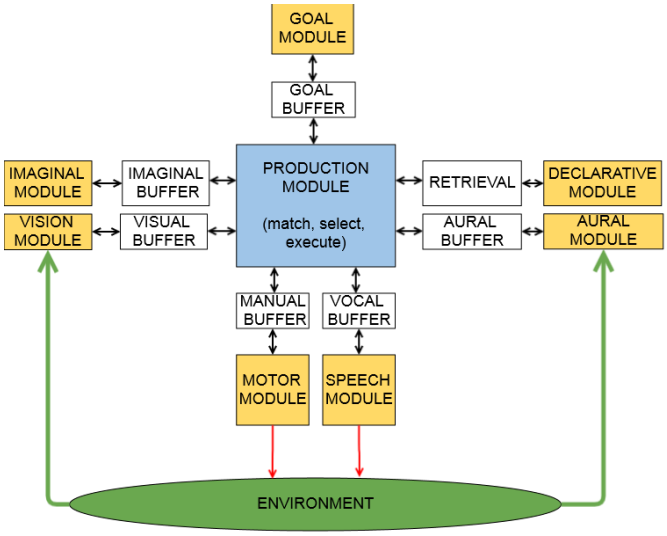


Figure 2. A schematic representation of the canonical ACT-R cognitive model.

ACT-R is formed as a knowledge model where the "chunks" are the elements of declarative knowledge in the ACT-R theory and are used to communicate information between modules through the buffers. A chunk is defined by its chunk type, that is described by its slots (here compared with properties), see Table I. Chunk types can be organized as a hierarchy of parent (SuperType)-child (SubType) relationships. The subtype will inherit all of the slots (properties) of the parent node(s).

Other models that take a similar symbolic approach to model human cognition include Soar [48], EPIC (Executive-Process/Interactive Control) [49], CLARION (Connectionist Learning with Adaptive Rule Induction On-line) [50], and others (for a detailed review see [51]). While these have been successful to varying degrees at modeling specific human cognitive task(s) performance, it is becoming evident that such models are intrinsically limited by their disconnection from the real world, in which humans (or robots) operate. A production based system is only as adaptive as its rule set allows given the inputs provided to it, which have generally been limited to "screen as eye" and "keyboard/mouse as hands" mappings. A new wave of thought surrounding the development of cognitive models is embracing the need for "embodied" cognition, improving the ability of the system to sense and act. One example of this is the ACT-R/E ("E" for "Embodied") framework, used as an operating system for mobile robotics developed by the American Naval Research Lab [52], depicted in Fig. 3.

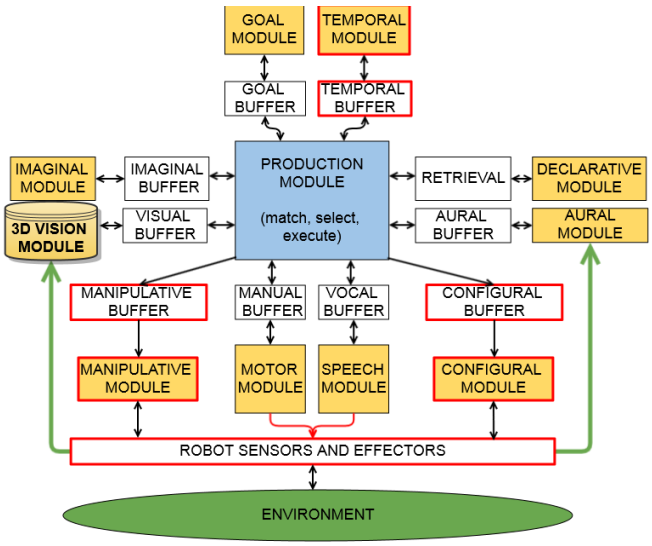


Figure 3. The "embodied" (Visual 3D) modifications introduced by Trafton et al. 2012. Additions in the ACT-R/E are highlighted in red.

The Object-Attribute-Relation (OAR) model of Wang, 2007 [53], specifies the elements of a cognitive model in the fashion of an ontology, the logical model of memory. In an attempt to formally describe the mechanism of human Long Term Memory (LTM), which he states is the "foundation of all forms of natural intelligence" (p. 66), Wang decomposes the construct into three elemental components – Objects, Attributes and Relations (OAR). This OAR model allows the computational specification of the human LTM formation and storage process, and is put forth as having sufficient explanatory power as to describe the "mental process and cognitive mechanisms of learning and knowledge representation" (p.72). This model has a strong parallel with the specification of knowledge in information processing ontologies. This parallel is direct, as described by the relations given in Table I.

TABLE I. COMPARISON OF MODEL TYPE CONSTRUCTS

OAR Model	Ontology Components	ACT-R ACT-R/E
Object(s) Attribute(s) Relation(s)	Class(es) Propert(ies) Relationship(s)	Chunk Type(s) Chunk Slot(s) Function(s)

A critical issue for any of these kinds of models is the relationship of their constructs to the environments, in which they are expected to provide foundations for action. The core notion of *embodiment* is to provide the heretofore functionally "disembodied" computational model with sensors and effectors that allow its direct interaction with the physical world. In such a way, the inherent limitation of human-defined input may be overcome. In addition to physical sensory perception and manipulative ability, a human may have access to a detailed semantic understanding of the surrounding world. In the quest to produce a non-human intelligent actor within a physical space, we must provide the actor with an understanding of underlying structures, i.e., specific denotations in the physical world.

VI. PROPOSED MODEL

In the CogOnto model, we propose a further augmentation of the cognitive models discussed above, providing the robot with detailed 3D schematic representations of objects that it encounters in real time, supported via task models, knowledge models and ontologies.

The CogOnto model is composed of five parts $\triangleq <S_i,C_i,A_i,O_i,R_i>$, where $i = 1..N$, and where S_i is a finite set of situations, C_i is a finite set of classes, A_i is a finite set of attributes for characterizing a class, O_i is a finite set of objects in a class, and R_i is a finite set of relationships among the objects. In the CogOnto model (Fig. 4), we consider the following features [54][56]:

- Situation: represents an interactive (i.e., dynamic) real world scenario.
- ConceptNet: is a network of class-to-class relationships applicable in a given situation.
- ObjectNet: an object is an instance of a class. ObjectNet is a network of object-to-object relationships.
- AttributeNet: is a network between properties of classes and objects.
- Relation: is a function associating concepts, classes, objects and attributes; e.g., a *robot* is *part-of* an *Intelligent Agent (IA)*, were the "part-of" relation connects two concepts. The relations (associations) may be modeled or created by an autonomous learning process.

These constructs are not defined in detail here, but unlike the other models are not limited to textual/linguistic meanings. The CogOnto model illustrated in Fig. 4 has four major functional elements that share information: 1) the ES-OE engine, 2) the eye tracking system interconnected with the ES-OE engine, 3) the OAR model functioning as the basis of the Cognitive System, and 4) the knowledge cloud, including external resources such as WordNet or Cyc. The latter is also called the Linked Open Data and may be used to illustrate the intelligent process for sharing and exposing information in machine readable form by using uniform resource identifiers based on Berners-Lee's [55][56] principles. These principles enable data communication guiding perception from procedural memory.

The knowledge system of the CogOnto model can be perceived as a storage system that accesses real world object

information and external semantic resource information via the existing knowledge cloud [57].

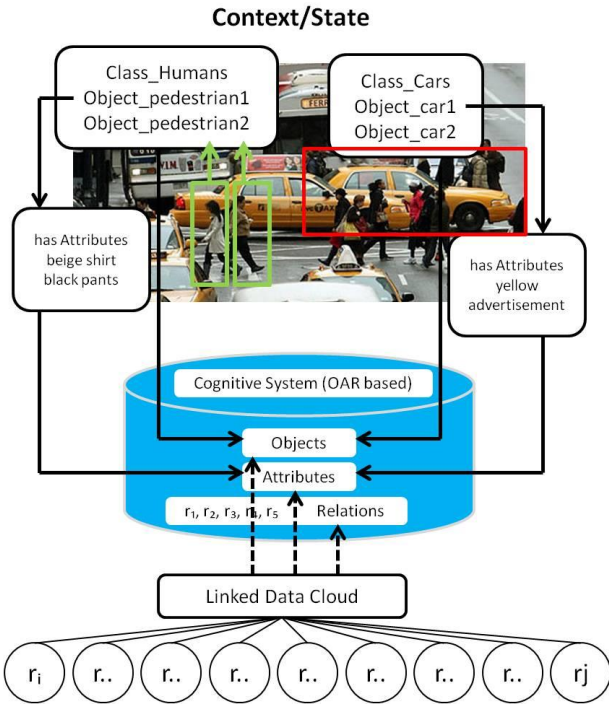


Figure 4. The CogOnto model and its operative states. The relations are build up for the current scene via object gaze tracking, and past stored scenes using a match function.

The knowledge system represents the integration of formal symbolic and free text descriptors of an object.

VII. INTEGRATING SEMANTIC WEB CONCEPTS, TECHNOLOGIES AND RESOURCES

CogOnto integrates its own knowledge resources with external resources accessible via the web. For example, WordNet is a lexical database where nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets). To recall an object, the 'synsets (WordNet 2.1)' [58] and the W3C [59] standard can be used at a text level, to describe what an object is when it is text-labeled. Ontologies can be expressed by using Semantic Web tools, e.g., Web Ontology Language (OWL) [60] and the Resource description framework Schema (RDFS) [61].

The OAR model, with its Object, Attribute and Relation parts, and the ontological framework, containing Class/Instance, Relationship and Properties, can be inter-mapped so the object world can be comprehended using existing resources and using the 3D information represented internally within an object model. The 3D object's internal structure and shape can either be structured as Free Form Geometry (FFG) with surfaces and curves, or as Polygonal Geometry (PG) with points, lines and faces. The objects can be extracted and exported into different file formats, such as, e.g., .obj files, .stl files. The .stl file format is a triangular representation of a 3D object, where each triangle is uniquely defined by its normal and three points representing its vertices. The format is native to the stereolithography Computer Aided Design (CAD) software created by 3D Systems (in this kind of format it is also possible to print the object out from a 3D printing machine).

The 3D object file contains different layers cognitively (form, volume, size, other descriptive attributes, etc.), supporting our senses and perception operating in parallel when performing allocated manipulation tasks. A human looking at an object can relate to the object both on a denotative- and on a connotative level. The denotative level is understood as a pure noun level without any cultural associations, nor any emotional or associative signifiers to the object, it is purely instrumental. The connotative layer is, on the other hand, the level of cultural and personal associations attached to an object with experience over time. Geometrical information within the 3D object can be represented using the X3D XML-based file format, an ISO standard for representing 3D computer graphics.

VIII. BEYOND ONTOLOGIES – COMPLEX RELATIONSHIPS, AND ALTERNATIVES TO HEIRARCHICAL DATA REPRESENTATION

As we move from relatively canonical data sets, for which the information processing ontology was designed (i.e., semantic relations within a particular knowledge base) to more complex relationships (such as ad-hoc physical relations), in which the hierarchical order is not nearly so explicit, or potentially non-existent, will the classical ontology suffice? Or alternately, will something more adaptive need to take its place? Because relationships in the physical world are multifaceted and multidirectional, it is useful to have a schema that can represent this interconnectedness. The key strength of an ontology is that it provides a concrete nomological environment, from which to operate within the chosen domain. Table II summarizes the traditional information processing ontology.

TABLE II. TRADITIONAL ONTOLOGY CHARACTERISTICS

- allows a common understanding of the structure of information
- enables reuse of domain knowledge
- makes domain assumptions explicit
- separates domain knowledge from operational knowledge
- defines a common vocabulary for researchers
- provides machine readable definitions of basic concepts and the relationships among them

However, there are instances (albeit few as of this writing), in which it is being recognized that the intrinsic limitations of the "ontology" such it is commonly

understood in 2014, (e.g., OWL-based) are sufficient as to demand a modification whereby the innate complexities of real-world phenomenon may be modeled. That is: complex, potentially non-hierarchical relationships.

For example, it has been noted in the field of chemical molecular informatics that while ontologies are able to represent tree-like structures, they are unable to represent cyclical or polycyclical structures [27]. Similarly, the difficulty in building classifications of nano-particles has led some researchers to begin to look into taxonomies based on "physical / chemical / clinical / toxic / spatial" characteristics of an object, supplemented by structural information, in order to account for shapes, forms and volumes [28]. Other examples of representing complex structural relations that stretch the boundaries of ontological representation include using Description Graph Logic Programs (DGLP) to represent objects with arbitrarily connected parts [29], and a hybrid formalism whereby the authors propose a "combination of monadic second order logic and ordinary OWL", where the two representations are bridged using a "heterogeneous logical connection framework" [30].

It is evident that the potential applications of a formalism such as the ontological method of information representation far outreach the initial conceptualizations of the language. While it may be possible to model 3D spatial information within the constraints of a hierarchical ontology, it is also to be considered that this notion, as well as applications such as those described above, may require the development of progressive, flexible alternatives, which capture the strengths of the ontology (i.e., the points from Table II), while managing to represent arbitrary or non-hierarchical relationships.

A. Cognitive Models and Ontologies

One information system where a non-hierarchical organization may be necessary, when attempting to map the internal structural relations, is the human brain. For more than half a century, researchers across many fields (e.g., Cognitive Psychology, Neuroscience, Cognitive Science) have been using models to posit and test hypothetical interpretations of how the human brain is structured. These range from the very simple (e.g., Baddely's working memory model, [31]) to complex neurological models (e.g., [32]), though no current model has even begun to approach the actual complexity of the human brain. On a neuronal level, and certainly even on a functional level such as between brain regions, this is a non-hierarchical system. It is once again remarkable that, at a superficial level, the development of ontologies draws a strong parallel with theoretical interpretations of how the human cognitive system might be structured (refer back to Table I). This relation is further discussed in Sennersten et al. [13].

In OAR (Object, Attribute, Relation), Wong [10] develops a model that most certainly shares conceptual roots with ontological knowledge representation. Likewise, parallels may be drawn with Anderson's ACT-R model [11] and Trafton's "embodied" version [32] ACT-R/E. In each model, *Objects* in the real world possess characteristics (i.e., *attributes*, or *properties*) and also *relations* with one another. If we can augment these heretofore largely semantic components with a functional representation of 3D space (e.g., at the 3 levels *Global*, *Local*, and *Personal*), we may have the fundaments of a system of Spatial Understanding for a robotic platform.

IX. CONCLUSION AND FUTURE WORK

The CogOnto model with support from the technological implementation of the eye tracker system with the ES-OE engine can represent cognitive relations that can be processed by a robot operating in a spatial world [62].

Formal knowledge structures within CogOnto face similar challenges to other knowledge representation formalisms, and this paper has shown isomorphism with a number of examples. However, the primary advance proposed is to use cloud-based resources that are not limited to formal representations to enhance the robustness of knowledge processing by the integration of similarity-based search. Those cloud-based resources may use text and images. But more interesting extensions for future work include new forms of cloud content, such as multi-spectral images, point clouds and behavior tracks. The main ongoing research challenge is to provide suitable similarity metrics for these data forms, integrating search results with formal structures, and developing methods for integrating them in unified search, or meta-search, results.

One of the few certainties regarding the immediate future is that robotic control technology will advance from systems that are coded for specific applications, to systems that are designed with an innate adaptability to unexpected environmental situations. This will require new methods of providing on-the-fly relational information to the robot, in order for it to gain an understanding of both its spatial position, and the position of other objects in the vicinity, their characteristics, and the ways that it can relate to them. A reworking of the traditional OWL-based ontology, with an eye for 3-dimensional spatial relations on 1) Global, 2) Local, and 3) Personal levels of specificity may be sufficient to this end.

It is also noted that as data sets become more complex, and especially as we begin to consider that most complex of biological control systems, the human cognitive system, it may very well become necessary to develop hybrid ontological-type systems of knowledge representation, which 1) encompass the full realm of advantages provided by the use of specific nomologial hierarchies, and 2) enable the encoding of arbitrary or non-hierarchical relationships.

The development knowledge-based systems that can account for abstract, non-hierarchical relations could potentially facilitate the next generation of spatially aware robotics applications.

REFERENCES

[1] M. Lochner, C. Sennersten, A. Morshed, and C. Lindley, "Modelling Spatial Understanding: Using knowledge representation to enable spatial awareness in a robotics platform," The 6th International Conference on Advanced Cognitive Technologies and Applications, 25-29th of May 2014, Venice, Italy.

[2] J. Kuffner, "Cloud Enabled Robots," Presentation, IEEE Humanoids conference, Nashville, Tenn. 2010. http://www.scribd.com/doc/47486324/Cloud-Enabled-Robots [retrieved: 2014.11.26]

[3] R. Khosla, M. Chu, R. Kachouie, K. Yamada, and T. Yamaguchi. "Embodying Care in Matilda: An Affective Communication Robot for the Elderly in Australia," In Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, pp. 295–304, 2012. http://dl.acm.org/citation.cfm?id=2110398. [retrieved: 2014.11.26]

[4] J. M. Lutin, A. L Kornhauser, and E. Lerner-Lam. "The Revolutionary Development of Self-Driving Vehicles and Implications for the Transportation Engineering Profession," Institute of Transportation Engineers, ITE Journal, vol. 83(7), July 2013, pp. 28.

[5] L. M. Hiatt, S. S. Khemlani, and J. G. Trafton, "An Explanatory Reasoning Framework for Embodied Agents," Biologically Inspired Cognitive Architectures, vol. 1, July 2012, pp. 23–31, doi:10.1016/j.bica.2012.03.001.

[6] A. Elfes, J. L. Hall, E. A. Kulczycki, D. S. Clouse, A. C. Morfopoulos, J. F. Montgomery, J. M. Cameron, A. Ansar, and R. J. Machuzak, "An Autonomy Architecture for Aerobot Exploration of the Saturnian Moon Titan," IEEE Aerospace and Electronic Systems Magazine, vol. 23(7), July 2008, pp. 1-9.

[7] K. K. Tahboub and S. N. Al-Din Munaf, "A Neuro-Fuzzy Reasoning System for Mobile Robot Navigation," JJMIE vol. 3(1), March 2009, pp. 77-88. http://pdf.aminer.org/000/361/105/a_neuro_fuzzy_approach_to_autonomous_navigation_for_mobile_robots.pdf. [retrieved: 2014.11.26]

[8] J. Fasola and M. Matarić, "Using Spatial Language to Guide and Instruct Robots in Household Environments," Refereed Workshop, AAAI Fall Symposium: Robots Learning Interactively from Human Teachers, Arlington, VA, Nov 2012. http://www.aaai.org/ocs/index.php/FSS/FSS12/paper/viewFile/5582/5880. [retrieved: 2014.11.26]

[9] C. Hudelot, J. Atif, and I. Bloch, "Fuzzy Spatial Relation Ontology for Image Interpretation," Fuzzy Sets and Systems, vol. 159(15), August 2008, pp. 1929–1951. doi:10.1016/j.fss.2008.02.011.

[10] G. Fu, C. B. Jones, and A. I. Abdelmoty, "Ontology-Based Spatial Query Expansion in Information Retrieval," On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE, Springer, 2005, pp. 1466–1482. http://link.springer.com/chapter/10.1007/11575801_33. [retrieved: 2014.11.26]

[11] Y. Wang, "The OAR model for knowledge representation," Proc. The 2006 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE'06), Ottawa, Canada, May 2006, pp. 1692-1699.

[12] J. R. Anderson, "ACT," American Psychological Association, vol. 51(4), 1995, pp. 355-365.

[13] M. Bosse, R. Zlot, and P. Flick, "Zebedee: Design of a Spring-Mounted 3-D Range Sensor with Application to Mobile Mapping," IEEE Transactions on Robotics, vol. 28(5), October 2012, pp. 1104–1119. doi:10.1109/TRO.2012.2200990.

[14] C. Sennersten, A. Morshed, M. Lochner, and C. Lindley, "Towards a cloud-based architecture for 3D object comprehension in cognitive robotics," The 6th International Conference on Advanced Cognitive Technologies and Applications, 25-29th of May 2014, Venice, Italy.

[15] C. A. Lindley. "Synthetic Intelligence: Beyond A.I. and Robotics," in Integral Biomathics: Tracing the Road to Reality, Simeonov, Plamen L.; Smith, Leslie S.; Ehresmann, Andrée C. (Eds.), Springer, 2012.

[16] T. R. Gruber, "A Translation Approach to Portable Ontology Specifications," Knowledge Acquisition, vol. 5(2), 1993, pp. 199–220.

[17] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, and Schneider, L, "Dolce: a descriptive ontology for linguistic and cognitive engineering," WonderWeb Project, Deliverable D17, vol. 2(1), 2003.

[18] C. Masolo, et al. "The WonderWeb Library of Foundational Ontologies," IST Project 2001 - 33052 WonderWeb: Intermediate Report, May 2003.

[19] J. A. Turner and A. R. Laird, "The cognitive paradigm ontology: design and application," Neuroinformatics, vol. 10(1), 2012, pp. 57-66.

[20] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to wordnet: An on-line lexical database," International journal of lexicography, vol. 3(4), 1990, pp. 235-244.

[21] C. Matuszek, J. Cabral, M. J. Witbrock, and J. DeOliveira, "An Introduction to the Syntax and Content of Cyc," AAAI Spring Symposium: Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering, March 2006, pp. 44-49.

[22] Laboratory for Applied Ontology (LOA) (ISTC-CNR) http://www.loa.istc.cnr.it/ [retrieved: 2014.11.26].

[23] D. B. Keator, et al. "A national human neuroimaging collaboratory enabled by the Biomedical Informatics Research Network (BIRN)," Information Technology in Biomedicine, IEEE Transactions on, vol. 12(2), 2008, pp. 162-172.

[24] A. R. Laird, J. J. Lancaster, and P. T. Fox, "Brainmap," Neuroinformatics, vol. 3(1), 2005, pp. 65-77.

[25] J. Dixon and O. Henlich, "Mobile Robot Navigation," Final Report, Information Systems Engineering, Imperial College, UK, 1997.

[26] Y. Kwoh, J. Huo, E. Jonckheere, and S. Hayati, "A Robot with Improved Absolute Positioning Accuracy for CT Guided Stereotactic Brain Surgery," IEEE Transactions on Biomedical Engineering, Feb 1988, vol. 35(2), pp. 153-160.

[27] J. Hastings, C. Batchelor, and M. Okada, "Shape Perception in Chemistry," Proceedings of the Second Interdisciplinary Workshop The Shape of Things (SHAPES 2013), Rio de Janeiro, Brazil, April 3-4, 2013, pp. 83-94. http://ceur-ws.org/Vol-1007/paper6.pdf. [retrieved: 2014.11.26].

[28] V. Maojo, et al. "Nanoinformatics: Developing New Computing Applications for Nanomedicine," Computing, vol. 94(6), March 7, 2012, pp. 521–539, doi:10.1007/s00607-012-0191-2.

[29] D. Magka, B. Motik, and I. Horrocks, "Modelling structured domains using description graphs and logic programming," Lecture Notes in Computer Science, vol. 7295, 2012, pp. 330-344, Department of Computer Science, University of Oxford, 2011.

[30] O. Kutz, J. Hastings, and T. Mossakowski. "Modelling Highly Symmetrical Molecules: Linking Ontologies and Graphs Artificial Intelligence: Methodology, Systems, and Applications," Lecture Notes in Computer Science, vol. 7557, chap. 11, pp. 103–111, Springer Berlin / Heidelberg, Berlin, Heidelberg, 2012.

[31] A. D. Baddeley and G. Hitch, "Working memory," In The psychology of learning and motivation: Advances in research and theory, vol. 8, G.H. Bower, Ed. New York: Academic Press, 1974, pp. 47–89.

[32] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," Nature Neuroscience, vol. 2(11), November 1999, pp. 1019–1025.

[33] G. Trafton, L. Hiatt, A. Harrison, F. Tanborello, S. Khemlani, and A. Schultz, "ACT-R/E: An Embodied Cognitive Architecture for Human-Robot Interaction," Journal of Human-Robot Interaction, vol. 2(1), March 2013, pp. 30–55. doi:10.5898/JHRI.2.1.Trafton.

[34] M. Bosse, R. Zlot, and P. Flick, "Zebedee: Design of a Spring-Mounted 3-D Range Sensor with Application to Mobile Mapping," IEEE Transactions on Robotics, vol. 28, no. 5, 2012.

[35] R. Zeibak and S. Filin, "Object extraction from Terrestrial Laser Scanning Data," TS 8E –Terrestrial Laser Scanning, Visualization and LIDAR, FIG Working Week 2009, Surveyors Key Role in Accelerated Development, Eilat, Israel.

[36] S. Westerberg and A. Shiriaev, "Virtual Environment-Based Teleoperation of Forestry Machines: Designing Future Interaction Methods," Journal of Human-Robot Interaction, vol. 2, no. 3, 2013.

[37] A. El Daher and S. Park, "Object Recognition and Classification from 3D Point Cloud," student project, 2006, http://cs229.stanford.edu/proj2006/ElDaherPark-ObjectRecognitionAndClassificationFrom3DPointClouds.pdf [retrieved: March 2013], Stanford Education at Stanford University, USA.

[38] C. Sennersten and C. Lindley, "Evaluation of Real-time Eye Gaze Logging by a 3D Game Engine," 12th IMEKO TC1 & TC7 Joint Symposium on Man Science and Measurement, Annecy, France, 2008.

[39] C. Sennersten, M. Castor, R. Gustavsson, and C.A. Lindley, "Decision Processes in Simulation-Based Training for ISAF Vehicle Patrols," NATO-OTAN, MP-HFM-202-17, 2010.

[40] P. Jerčić, et al., "A Serious Game Using Physiological Interfaces for Emotion Regulation Training In The Context of Financial Decision Making," ECIS 2012 Proceedings. AIS Electronic Library (AISeL), 2012.

[41] H. Cederholm, O. Hillborn, C. Lindley, C. Sennersten, and J. Eriksson, "The Aiming Game: Using a Game with Biofeedback for Training in Emotion Regulation," 5th Digital Games Research Association (DIGRA) Conference THINK DESIGN PLAY 2011, Utrecht, Netherlands.

[42] J. R. Flanagan, G. Rotman, A. F. Reichelt, and R. S. Johansson, "The role of observers' gaze behavior when watching object manipulation tasks: predicting and evaluating the consequences of action," Philosophical Transactions of the Royal Society –B: Biological Sciences, 2013, UK.

[43] Oxford English Dictionary entry: http://oxforddictionaries.com/definition/english/knowledge, [retrieved: 2014.11.26].

[44] J. R. Anderson, "ACT," American Psychological Association, Vol. 51, No.4, 1995, p. 355-365.

[45] D. Kieras and D. Meyer, "An overview of the EPIC Architecture for Cognition and Performance with Application to Human-Computer Interaction," University of Michigan, EPIC report No. 5 (TR-95/ONR-EPIC-5), 1995, DTIC Document 1995, 43 pages.

[46] D. P. Benjamin, D. Lonsdale, D. Lyons, and S. Patel, "Using Cognitive Semantics to Integrate Perception and Motion in a Behavior-Based Robot," In Learning and Adaptive Behaviors for Robotic Systems, 2008, LAB-RS'08. ECSIS Symposium on, pp. 77–82. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4599431 [retrieved: 2014.11.26].

[47] E. Hutchins, Cognition in the Wild. Chapter 9 – Cultural Cognition, 1996, pp. 353-374.

[48] J. Laird, A. Newell, and P. Rosenbloom, "SOAR: An Architecture for General Intelligence," Technical report AIP-9, University of Michigan, Carnegie-Mellon University, Stanford University, Artificial Intelligence 33, 1987, pp. 1–63, DTIC Document 1988, 63 pages.

[49] D. Kieras and D. Meyer, "The EPIC architecture for modeling human information-processing and performance: A brief introduction," EPIC Report No.1 (TR-94/ONR-EPIC-1), University of Michigan, 1994, DTIC Document 1994, 43 pages.

[50] R. Sun, "The CLARION Cognitive Architecture: Extending Cognitive Modeling to Social Simulation," Cognition and

Multi-Agent Interaction, Oct. 2004, Cambridge University Press, New York, 2006.

[51] D. Vernon, G. Metta, and G. Sandini, "A Survey of Artificial Cognitive Systems: Implications for the Autonomous Development of Mental Capabilities in Computational Agents," IEEE Transactions on Evolutionary Computation 11 (2), April 2007, pp. 151–180, doi:10.1109/TEVC.2006.890274.

[52] G. Trafton, et al., "ACT-R/E: An Embodied Cognitive Architecture for Human-Robot Interaction," Journal of Human-Robot Interaction, vol.2, No.1, 2013, pp. 30-55.

[53] Y. Wang, "The OAR model for knowledge representation," Proc. The 2006 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE'06), Ottawa, Canada, pp.1692-1699.

[54] A. Oltramari and C. Lebiere, "Extending Cognitive Architechtures with Semantic Resources," Department of Psychology, Artificial General Intelligence Lecture Notes in Computer Science, Vol. 6830, 2011, pp. 222-231.

[55] T. Berners-Lee, "Linked data-the story so far," International Journal on Semantic Web and Information Systems, 5(3), 2009, pp. 1-22.

[56] Linked_Data_Design_Issue. http://www.w3.org/DesignIssues/LinkedData.html [retrieved: 2014.11.26].

[57] C. D'este et al., "Sustainability, Scalability, and Sensor Activity with Cloud Robotics," Proceedings of Australasian Conference on Robotics and Automation, 2-3 Dec 2013, University of New South Wales, Sydney, Australia.

[58] WordNet: http://wordnetweb.princeton.edu/perl/webwn [retrieved: 2014.11.26].

[59] World Wide Web Consortium is the main international standards organisation for the World Wide Web.

[60] OWL W3C: http://www.w3.org/TR/owl-features/ [retrieved: 2014.11.26].

[61] RDFS W3C: http://www.w3.org/TR/rdf-schema/ [retrieved: 2014.11.26].

Combining Cognitive ACT-R Models with Usability Testing Reveals Users Mental Model while Shopping with a Smartphone Application

Sabine Prezenski
sabine.prezenski@tu-berlin.de

Nele Russwinkel
nele.russwinkel@tu-berlin.de

Dep. of cognitive Modeling in dynamic HMS
TU Berlin
Berlin, Germany

***Abstract*—The usability of two different versions of a smartphone shopping list application for Android is evaluated via user tests and cognitive modeling. The mobile application enables users to compose a shopping list by selecting items out of different stores and product categories. The two versions of the linear hierarchical application differ in menu depth. Two empirical studies compare novice and expert product search time. The first study focuses on efficiency, suitability for learning, mental load and mental models. The second study supplements the findings of the first study and investigates varying expectations between products. An ACT-R based cognitive modeling approach provides in depth explanations for the effects found in the empirical study. The study shows that for expert users, product search with a 3 layer or a 2 layer version is equally efficient, due to the same amount of mental load. Expert and novice user rely on different strategies when searching for items- novice users need to access their general knowledge frequently, experts use their mental model of pathways leading to the items. The suitability of the mental model of users, explains why version updates that introduce a new layer produce longer product search times - and those reducing the number of layers do not.**

Keywords-cognitive modeling; ACT-R; usability; smartphone; application; menu; mental load; mental model.

I. INTRODUCTION

Nowadays, life without mobile applications and smartphones is hard to imagine. New evaluation methods for mobile applications are needed [1] because the market for is growing rapidly [2]. For an application to be successful, high usability is compulsory. Conventional usability testing is time and money consuming. Therefore, a pressing question is how the usability of applications can be guaranteed, without costs exploding. On the long run, cognitive models can serve as a substitute for usability testing. The following paper is a step towards this goal.

The current paper investigates menu depth in a real-world setting with a new smartphone application. In contrast to this, most studies concerning menu depth use artificial labels and tasks in a laboratory setting.

Our work demonstrates that the most important factor for menu design is not reducing the number of clicks, but users' mental processes. An important finding is that the best number of levels of menu hierarchy may differ from case to case, but is one that maintains users' mental load to a minimum. Well designed applications address users' mental models of these applications. The fact that mental models of users are not static, but evolve as they develop from novices to experts is a further topic of this paper. The learning processes while handling a new application is studied.

In the current work, cognitive modeling is used to explore users' mental processes. We will demonstrate how ACT-R based cognitive modeling explains results obtained in empirical usability studies. It is shown, that cognitive modeling with ACT-R has the potential to replace traditional user tests to a certain extent, but also help to understand the underlying mental mechanisms in this kind of human-machine interaction.

The empirical part consists of two studies on two different versions of a shopping list application. The Android application allows users to select products out of a categorized hierarchical list or via an alphabetical product overview. With the application, users can compose a shopping list. The two versions differ in menu depth. Although both studies concern the same application, design and purpose of the studies differ.

The first study allows a conclusion on the overall usability of the application, due to the fact that all navigating possibilities, the app provides, are allowed. The sample size of the study permits statistical testing to compare the versions, too. In the first study, an ACT-R modeling approach is introduced. The second study supplements the first study. It restricts functionality of the application in order to substantiate the model assumptions about mental models of users from the first study. Furthermore, the second study enables to conduct learning curves and to investigate different expectations on product affiliated categories.

In both studies, users repeatedly search for the same products. Therefore, novice and expert users can be studied and the suitability for learning of the application can be evaluated.

The modeling part further addresses how mental models of novice and expert users develop as users become more experienced with an application. It also investigates how version updates of software challenge users' mental model. This study also unfolds the relationship between menu hierarchy and cognitive load.

II. THEORY

A. Usability

Standard ISO 9241-11 specifies usability as effectiveness, efficiency and satisfaction. General ergonomic principles for the design of dialogues between humans and information systems are specified in standard ISO-9241-110, which outlines seven criteria (suitability for the task, suitability for learning, suitability for individualization, conformity with user expectations, self descriptiveness, controllability, and error tolerance). Nielson`s Usability Heuristics are another way of specifying usability; they describe ten general principles for interaction design, for example that consistency and standards should be applied for successful applications [3].

1) Measurement of Usability: There are various methods to judge the usability of mobile applications; user focused assessment makes an important distinction between expert reviews and user data. A more engineering centered approach is, e.g., the method of pattern matching [4], which allows designers to assess certain usability problems, without interacting with users. There are different approaches to evaluate user data; either via qualitative methods (e.g., think aloud protocol), questionnaires or user tests. Particularly, information about subjective satisfaction can only be obtained with qualitative measurements, e.g., questionnaires or interviews. Nevertheless, high correlations between subjective satisfaction and quantitative measurements of usability are expectable [5] [6]. Therefore, assuring effectiveness and efficiency is important for achieving subjective satisfaction. Quantitative user testing allows assessment of a wide range of usability criteria; e.g., task completion time as a measure for efficiency, the number of successful task completions in a given time as a measure for effectiveness. The number and kind of mistakes give information about suitability of the application for the task, about conformity with user expectations, about self descriptiveness, controllability and error tolerance. Suitability for learning can be measured via comparison of several runs. Furthermore, contrasting inexperienced users (so called novice user) and very experienced users (experts) provides an interesting insight into the question if experience with an application provides a benefit for users. A reduction in time on task is expected for expert users, since they have developed a mental plan for the handling of an application. One should be aware of the fact, that not only performance in terms of time on task improves with experience, but that the structure of individuals knowledge (their mental model) changes as well [7].

The problems with user tests and questionnaires are similar to those of general psychological test; various testing aspects (such as reactance, conformity and other motivational issues) influence the outcome. Besides psychological testing effects, user tests are expensive and time consuming. Individuals need to be recruited and tested. Therefore, alternative methods that do not require user testing would be of value. Furthermore, methods that state precise concepts, which can then be transferred to other applications, are eligible. Cognitive Modeling fulfills the requirements mentioned above and is further a method to assess usability. Cognitive modeling is a helpful approach to learn more about cognitive processes during the interaction with applications. In addition, cognitive modeling offers the opportunity to explore the structure of users' knowledge. In the future, predictive cognitive models can serve as a substitute for user tests.

User tests can assess the most important aspects of usability, such as effectiveness and suitability for learning. Quantitative measurements can be replaced by cognitive models. In addition, cognitive models offer explanations about mental processes influencing usability, a benefit that goes beyond the scope of simple user tests.

2) Usability and Smartphones: In the field of mobile applications, special challenges for usability testing exist. Especially, aspects of mobile context, limited connectivity, small and varying display size and aspects concerning data entry methods should be accounted for [8]. In a review on different studies on usability of mobile applications, Harrison et al. [9] stress the importance of mental load of applications for successful usage.

Mental load is defined as the mental cost required fulfilling a task [10]. Mental (or cognitive) load is a multidimensional concept, with subjective, objective and psycho-physiological components and therefore difficult to measure [11]. The PACMAD (People At the Centre of Mobile Application Development) usability model for mobile devices includes mental load into the ISO definition and further incorporates the user, the task and the (more mobile) context [10].

It is highly questionable if mental load, a multidimensional and crucial concept for mobile usability, is assessable with user test. User tests can assess the most important aspects of usability, such as effectiveness and suitability for learning. Quantitative measurements can be replaced by cognitive models. In addition, cognitive models offer explanations about mental processes influencing usability, a benefit that goes beyond the scope of simple user tests.

B. Modeling and Usability

It is stated that mental load is impossible to asses via heuristics or standards [9]. Hence, a different approach is needed. On the other hand, assessing cognitive load with cognitive models is possible and already carried out [11] [12].

CogTool [13] and MeMo [14] are tools that allow user modeling of smartphone applications and websites and provide insights about usability problems.

CogTool is a user interface prototyping tool, which produces a simplified version of ACT-R [15] code; it is based on keystroke-level modeling [13]. KLM divides tasks as into different kind of actions (e.g., keystrokes, pointing) and mental processes, which are represented through mental operators [16]. A specific amount of time is assumed for

each action and each mental operator. Total task time is composed of the sum of these. In order to produce a cognitive model with CogTool, one has to manually click together a storyboard. The model then runs along the pathway as described in the storyboard. CogTool predicts how long a skilled user will take for the specified task [13]. CogTool has limitations, for example, it is not possible for the model to explore the interface since the model only runs along the ideal-pathway (as defined by the storyboard). As a result of this, information about potential user errors or the influence of workload cannot be achieved. Furthermore, information about learning or the difference between expert and novice users cannot be uncovered using CogTool.

MeMo is a Usability Workbench for Rapid Product Development, which can simulate user interactions with the system [14]. On the basis of tasks, possible solution pathways are searched by the model and deviations from these pathways are then generated; different user groups (e.g., elderly users, novice users) are taken under consideration [14], which is clearly an advantage of MeMo over CogTool. Another advantage of MeMo is that the model can produce errors, just as real users would do. A distinct disadvantage of MeMo is that it is not a cognitive modeling tool- important concepts about human cognition such as learning are not implemented. Therefore, the validity of the conclusions attained with MeMo is questionable.

Besides the introduced ready-to-use tools, there are numerous modeling approaches that uncover how different cognitive aspect or design factors affect usability. These modeling approaches attempt to describe and predict coherence between usability influencing aspects. Results obtained from the modeling approaches introduced below, can be used to derive general advice for designers.

1) Mathematical Models: Mathematical models are developed to predict measurements of usability, such as selection time as a function of external factors [16]. For example, the relation between item position and target search time in a linear menu can be described as a predictive mathematical model [17]. Other mathematical models focus on how factors such as menu size, target position and practice influence usability factors [16]. Such numerical models provide straightforward advice to designers. But on the downside, they are not helpful in identifying why these relations exist and give no information about learning and workload influence. One does not gain insight on how or why ongoing cognitive processes influence usability.

2) Cognitive Models: Therefore, to reveal the causes of differences in usability performance measurements, cognitive models should be consulted. Just as mathematical models, they provide numerical predictions. But cognitive models simulate the interaction with an application in the way users would interact with the application. Specific cognitive processes such as attention, perception and memory are incorporated in these kinds of models and can serve as an explanation for differences in usability findings. The best way to build scientifically grounded cognitive models is to use cognitive architectures.

a) Cognitive Architectures and ACT-R: Cognitive architectures offer a computable platform that represents well established theories about human information processing. With cognitive architectures, it is possible to simulate cognitive mechanisms and structures such as visual perception or memory retrieval. EPIC [18] and ACT-R [15] are two architectures used for modeling aspects of human computer interaction. This paper focuses on an ACT-R model of user interaction. To understand the modelling approach descirbed later it is helpful to know about the core mechanisms of ACT-R [15]. ACT-R is a hybrid architecture, which means that it has symbolic (knowledge representations such as chunks and rules called productions) and sub symbolic components (activation of chunks and utility of productions). The symbolic part consists of different modules and their interfaces (called buffers), with which these modules communicate with the production system. Only one element can be stored in each buffer at a given time. Similar to the brain, ACT-R distinguishes different areas called modules, which process certain classes of information. For instance, the declarative memory module, can store information in units called chunks. These chunks can be retrieved, which means that a chunk, which matches the given criteria is put into the according buffer and can be processed further by the production system. New chunks are constructed in the imaginal module. Other modules process visual or auditory information. There are also output modules such as vocal and motor modules. These are just some of the available modules. Furthermore, the sub symbolic components of the architecture are important. If some chunks are retrieved and used more often than other chunks, these chunks are given a higher activation level. This activation level determines how quickly a chunk can be retrieved or if it can be retrieved at all. Information that is not often used will decay over time and at some level will be forgotten and hence cannot be retrieved. The structure of chunks is characterized by different slots (or attributes) that can be filled with information. Category membership is represented in slots; this allows building semantic networks. Furthermore, new chunks can be learned during a task. The production system persists of rules defined by an “if” and “then” part. If the cognitive system with its modules and chunks in the buffers meet the conditions of the rule, the rule can be selected. In this case, the action part is executed. The production systems enables to initiatechanges to the chunks or to send requests to the modules (e.g., to the motor module “press mouse button”). If particular rules are more useful than others they receive a higher utility level and will be preferred to others. Also, reward can be given to productions if they lead to a goal, which also influences the utility level. It is possible to enable a process called production compilation. Here productions can be combined if they precede each other often or if identical chunks are

frequently retrieved in similar situations. This way the model will become faster just as human behavior improves as a task is done multiple times.

b) ACT-Droid: In the modeling approach presented, a new tool, ACT-Droid [19] is implemented, which has outstanding benefits to other approaches. Instead of replicating mobile applications or websites as mock-ups or reprogramming aspects of the application for the model, ACT-Droid allows developing user models that directly interact with Android smartphone applications. The ACT-Droid tool enables a direct connection of the cognitive architecture with an Android smartphone application via TCP/IP protocol. With this tool the modeling process becomes more convenient and much faster. In general one has to define a simlple interface version of the application in Lisp, with which the model can then interact with. When using ACT-Droid the ACT-R model can directly interact with the original Android application. The user model can interact with buttons and perceive changes on the interface.

The tool has many advantages for the modeler. First of all, no mock-up version of the app or possible pathways need to be created, which saves a lot of time, compared to CogTool or MeMo. Secondly, models interacting with the application, can implement the full possibility and functions of the ACT-R architectures, which allows investigating a great number of different aspects of how applications affects human information processing and individual differences (e.g,. memory, learning, experience or age). Thirdly, with this modeling approach processing time as well as different kind of user mistakes can be evaluated.

Main requirement for the usage of our approach are skills in modeling with ACT-R. The modeler just needs to know how to write (or change) productions and have rudimentary knowledge of the sub symbolic part of ACT-R. No lisp-programming is needed. Thus, ACT-Droid makes modeling with ACT-R much less complicated and more straightforward.

c) Modeling of Smartphone Applications: Applications for smartphones are small programs. Most applications are very specific for their field and are hence built to solve limited tasks. The limited scope of applications and the fact that successful applications have a simple and consistent design, make them profound for cognitive modeling. Developing a user model able to interact with the application is an accomplishable modelling task and can help uncover difficulties in the application that negatively influence usability. Some factors influencing the usability of smartphone applications, especially mental load or aspects concerning mental models are especially eligible to be evaluated with cognitive models.

Building up a mental model of an app the user normally orients oneself towards the menue structure.

C. Menus

An important research question concerning usability is how a menu structure should be designed in order to offer the best opportunity for navigating an application [20]–[23]. Menus help users find the right information. Different types of menus exist, e.g., square menus, pie menus, linear menus, hierarchical menus [23]. This paper focuses on linear hierarchical menus. Research on menu structures and design has revealed many important factors contributing to successful usability of menus. The following findings are derived from strongly controlled laboratory studies or studies dealing with either desktop menus or website menus. Consequently, it is questionable if these findings can be transferred to real-life smartphone applications.

Zhang states that clear and consistent labeling, predictability, a minimum of interaction steps, but also the avoidance of long list in a menu structure are import factors, contributing to the usability of menus [8].

Nilsons [17] identifies important factors that influence item selection time, such as menu length, item placemen and menu organization. Shorter menus are beneficial; e.g., users are faster in selecting and searching items from shorter than from longer menus [24]. When it comes to menu organization, organization of items in a menu is more beneficial than random placement of items [25]. Semantic and alphabetic organizations are two typical ways of organizing items. The target position of an item on a menu has a strong influence on item search time. Targets positioned on the top of a menu are found faster than those positioned in the middle [26]. Targets positioned on the bottom of the menu are likewise easier to be found [16]. The more users are familiar with a menu, the less time they take for finding an item, this is known as practice effect [20] [27]. If the target item is included in the menu, scanning is quicker, than if the target is not included in the menu [16]. Target items are found faster, if the item label is strongly associated to the target item [21].

a) Menu Hierarchies: Menu hierarchies are another factor that influences the usability. Lee and MacGregor [28] point out that two main factors influence search time for an item in a hierarchical menu; the number of pages (or levels) that have to be accessed and the time required to select alternatives from pages. The required time is directly dependent on the number of alternatives per page. For smartphones, finding an adequate depth and breath for the menu hierarchy is especially important. The small display size and scrolling time, make it even more important to provide users with and easy accessible and transparent menu. Some design experts recommend, that menus of mobile phones should rather be narrow and deep than shallow and broad [29]. Others state that adding more hierarchy levels (especially for menus with more items) is advisable, but that hierarchies with more than three levels should be avoided [30]. In General, findings in the literature are conflicting [31]. Another aspect influencing hierarchies

is scrolling. Cockburn and Gutwin propose a mathematical model linking the relationship between menu- scrolling and hierarchy on desktop computers [31]. When investigating an adequate ration between menu depth and items, instead of discussing the numbers of items or menu level, one can also focus on the question if users mental model matches the model of such a menu [23] [30]. Since measuring mental models is doubtful with classical methods and ACT-Droid provides the possibility to model application, this essential topic of menu hierarchies should be studied with cognitive modeling.

b) Menu Structures: Currently, most modeling studies on menu design focus mainly on visual processing [16] [24] [26] [32] [33] for evaluating menu design and usability aspects. According to an EPIC model from 1997 [32], eye movement pattern are a 50/50 mixture of sequential top-to-bottom and randomly searching. When serial top-to-bottom search is executed, the users' eye moves down the menu with constant distance in each saccade. Because of parallel examination of multiple items, the eyes regularly pass the target item by one saccade. Motor movements do not occur before the target is located. An ACT-R model from the same year [33] on the other hand predicts, that eye movements are exclusively top-to-bottom, and that the distance of each saccade varies. The eyes never pass the target items and that motor movement follows the saccades and happens before the target is located. Succeeding studies [24] [26] that used an eye tracking experiment and an ACT-R / PM model found that the first eye fixation is almost always towards the top of a menu and that most of the time visual search is top-to bottom. Search is rarely random. Some items are skipped and then backtracked. These models, with a focus of visual encoding, provide explanations for effects, such as the preferred item position. Mathematical models have a strong focus on different visual search strategies as well. Serial search and direcfted search are two modeled visual search strategies [16]. Serial search labels, top-to-bottom search and directed search describe search as determined by focusing to the assumed location of the target. Mathematical models also suggest that visual search of novice and expert users follows different functions [20]. These models describe the fact that as learning proceeds performance increases, which is explained through remembering the position of visual items [20]. As a conclusion, most modeling studies dealing with aspects of menu design, focus on visual and motor processing and few studies compare expert and novice performance [20]. Even though empirical studies indicate that menu-structures should incorporate the mental model of potential users and claim that mental models changes as users become more familiar with the menu structure [25] [21], as far as the authors are aware, no cognitive modeling studies focusing on this exists. This paper will present how cognitive model can address user's mental models of menu structure of applications.

III. METHODS

A. Purpose of this study

The study concept is designed to investigate menu hierarchies of an application. Two versions of an application, differing in menu depth, are compared using concepts derived from cognitive modeling. One version has two subcategories with the disadvantages of more required clicks; the other has only one level of sub-categories and therefore requires fewer clicks. In this paper, we develop modeling concepts for different aspects of usability. Aspects, such as efficiency, suitability for learning and the development of mental models are measured. A combination of empirical data and cognitive modeling approaches is presented.

We propose that, when it comes to menu structures, it is important that a menu should be designed to fit the cognitive capabilities of humans. If designers focus on reducing clicks, they might miss the turning point, that less clicks are associated with more cognitive load (e.g., memory load). In general, care should be taken in the design of applications, so that the principles of optimal human information processing are met. It is commonly agreed, that human knowledge is represented in form of a semantic network [34]. Within this network, categories are associated with subcategories and retrieval of subcategories succeeds best and faster when the category representations are addressed. In the study we will assess how well an application meets the mental model that users have about the application.

Novice (first interaction with a new version) and expert (second interaction with a new version) behavior will be compared, so information about the evolving mental model of users can be obtained. Furthermore, the suitability for learning can also be measured. Most studies investigating menu designs are strongly controlled and hence artificial laboratory studies [16] [26]. Quite the contrary is the case for the current study, which is conducted with a smartphone application. Furthermore, a very realistic task is utilized. Test persons are asked to select products from a shopping list application which provides them with a ready to use shopping list. This paper presents a combination of an empirical study of the usability of an Android shopping list application with cognitive modeling approaches. Cognitive modeling of the user behavior incorporates the full ACT-R architecture. Although visual processing of menus is modeled, this study focuses on the development of an adequate mental model and learning processes as users would do it.

B. The Application

Both versions of shopping list application are designed for Android. The application allows users to select products out of either an alphabetically ordered list or via categorical search (see Fig. 1). The chosen products are then added to a list. The difference between the two versions is menu depth: The three layer version (*3L*) has one more menu level than the two layer version (*2L*). The first page of the application is the same for both versions: three buttons are presented: "overview", "shops" and "my list". For both versions, after

selecting "overview" the list of the alphabet appears. Three or two letters are always grouped together on one button, e.g., "ABC", "DEF"…. Selecting one of those buttons then results in an alphabetical ordered list of the products. A click on a small checkbox in the right of the product selects it. If users click on shops, the *categorical pathway* is accessed. For both versions, clicking on shops results in a list of seven shops (bakery, drugstore, deli, greengrocer, beverage store, stationery, and corner shop). Each of these shops is represented by a button. For the *2L version*, selecting one of the shops results in an alphabetical ordered list of the products available in that particular shop. For example, by clicking on greengrocers all items that can be found in a greengrocers store are presented (apples, bananas, blueberries, cherries, etc.) and are selected by a click on the checkbox. For the *3L* version, the shops have seven subcategories, each. For example, when selecting greengrocers, one is presented with the subcategories exotic fruits, domestic fruits, tuber vegetables, herbs, seeds and nuts, mushrooms and salads. When selecting a subcategory, a list of products that can be found under this subcategory, appears and can be selected via the checkbox. For both versions, selecting "My List" from page one results in a shopping list, which comprises the selected products plus information about the store, in which the products are available.

Figure 1. Different product pathways for alcohol free beer. The orange path is the alphabetical pathway. The green and blue paths are the categorical pathway. The green pathway is the pathway of the *3L* app, the blue of the *2L* app.

C. Procedure

The first study is designed in order to investigate if the user interaction with the two versions differs on a statistical significant level. It further allows a conclusion on the overall usability of the application- namely on efficiency, effectiveness and suitability for learning. To ensure conditions close to real-life interaction, all navigating possibilities the application provides are allowed. An ACT-R modeling approach concerning the evolvement of user's mental model and the influence of cognitive load is also presented. The second study supplements the first study. The functionality of the application is restricted, only the categorical pathway is allowed for product search. This restriction is necessary for two reasons. First, it substantiates assumptions about mental models of users derived from the first study. Second, the restriction allows investigating learning curves. With help of the learning curves differences between products can be uncovered. In both studies participants were asked to find products, using both versions of the shopping list application. The application was presented on a Google Nexus 4 smartphone, running with Android 4.1.2. Each product was read to the participant by the experimenter and then the participant was asked to find the product and select it. In the first study participants were free to choose the pathway, which led them to the products. They were instructed to use the different possibilities the app provided, so they could either find the products via the *alphabetical* or via the *categorical* pathway. In the second study, participants were asked to select products merely using the *categorical* pathway. 26 student participants (12 male and 14 female, age_{mean}= 23) participated in the first study and 17 student participants (6 male and 11 female, age_{mean}= 26) participated in the second study. After receiving standardized oral instructions participants were instructed to select a list of products. For each trial a product was read to participants by the investigator and participants had to find the product. After selecting a product, participants were asked to return to the first page and then the next trial started.

TABLE I. DESIGN OF STUDY 1

order of versions	*3L* (new)	*3L* (expert)	*2L* (new)	*2L* (expert)
***3L* first, *2L* second**	Block 1	Block 2	Block 3	Block 4
***2L* first, *3L* second**	Block 3	Block 4	Block 1	Block 2

Enforcing the participants to always return to the first page (e.g., starting point) was necessary for reasons of experimental control. After selecting eight or nine products, participants were asked to read the shopping list (in order to assure learning of the store categories). For the next block, the items were identical but presented in a different sequence. After completing the second block, the investigator presented the participant the other version and the two blocks of trials were repeated. For the first study half of the participants first worked with the *2L version* and the other participants began with the *3L version* (see Table I). In the second study all participants first worked with the *2L version* and then switched to the *3L version* (see Table V).

IV. RESULTS

A. Study 1

1) Hypothes:

a) The first study investigated how product search time is influenced by menu depth, expertise and version specific expectancies, e.g., users' mental model of the respective application. Product search time is an indication of efficiency.

b) The main difference between both *versions* of the application is menu depth. We expected overall product search time to be longer for the three layer version than for the shallower two layer version.

c) As participants become more familiar with the application, we expected product search time to decrease as experience increases. Otherwise, the application would not be *suitable for learning*.

d) Finally, we were interested in how *version specific expectations*, gained with one version of the application can influences performance in product search with the other application. We propose that learning transfers from one version to the other will occur.

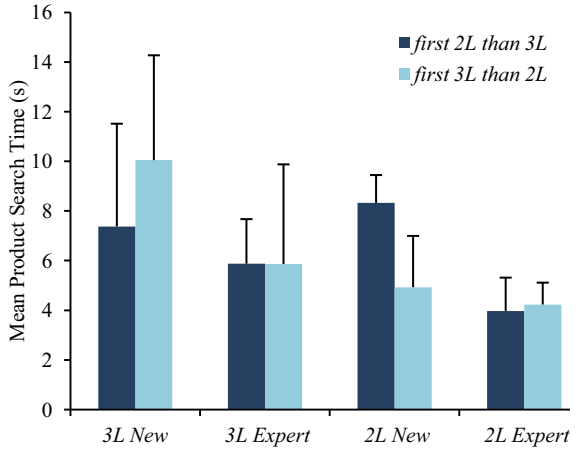


Figure 2. Mean trial time of study 1.

2) Descriptive results: All products could be found with both versions by all groups of participants, therefore, effectiveness of the application is given. Fig. 2 shows the mean trial time and standard deviations for the different conditions. The green bars represent the group *2L first*, *3L second* and the blue bars the group *3L first*, *2L second*.

For participants of group *3L first*, *2L* second, the mean trial time of block 1 (*3L* new) is 10.048 seconds and decreases approximately 4 seconds for block 2 (*3L* expert) (mean trial time 5.861 seconds). After switching to the *2L* version (*2L new*) time decrease to 4.928 seconds (block 3) and reaches 4.229 seconds for *2L expert* (block 4). For participants of group *2L first*, *3L second* a trial in the first block (*2L new*) has a mean duration of 8.322 seconds and a trial in the second block (*2L expert*) a mean duration of 3.971 seconds. After participants switch to *3L* version (block 3, *3L new*) time increase to 7.376 seconds and decreases again to 5.875 seconds (block 4, *3L expert*).

3) Statistical analysis and results: To investigate how product search time is influenced by menu depth, expertise and version specific expectancies a 2x2x2 ANOVA was conducted.

The following factors were considered: the factor *order of the versions* with the two steps *"3L first, 2L second"* and *"first 2L than 3L";* the repeated measurement factor *version* with the two steps *"3L version"* and *"2L version"* and the repeated measurement factor *expertise* with the two steps *"new"* and *"expert"*.

The overall result of Levene test for sphericity was not significant therefore, the overall distribution of the error variance is equal in all groups and an ANOVA could be performed.

The ANOVA reveals a significant main effect of factor *version* with $F(1,24)=12.527$, $p<0.005$ and a medium effect size (partial $\eta 2=0.343$). Descriptive results indicate that the *2L version* is overall faster than the *3L version*, indicating that shallower menu depth, results in less search time. This effect is labeled *version effect.* Another significant main effect is found for the factor *expertise* $F(1,24)=29.625$, $p<0.001$ and a medium to large effect size (partial $\eta 2 =0.552$). Descriptive results show, that performance in the new conditions is slower than in the expert conditions, which is a clear indication that learning occurs. This effect is labeled *experience effect.*

The interaction between *version* and *order of the versions* is also significant $F(1,24) =7.076$, $p<0.05$, with a medium effect size (partial $\eta 2=0.228$). The interaction between *version, novelty* and *order of the versions* is further significant, with $F(1,24) =13.661$, $p<0.001$ and a medium effect size (partial $\eta 2=0.363$).

TABLE II. DESCRIPTIVE STATISTICS OF STUDY 1

		Mean	Std. Deviation	N
3L new	*2L* first, *3L* second	10.048	4.137	13
	3L first, *2L* second	7.376	4.220	13
	total	8.712	4.315	26
3L expert	*2L* first, *3L* second	5.861	1.793	13
	3L first, *2L* second	5.875	4.013	13
	total	5.868	3.045	26
2L new	*2L* first, *3L* second	4.928	1.122	13
	3L first, *2L* second	8.322	2.063	13
	total	6.625	2.375	26
2L expert	*2L* first, *3L* second	4.229	1.340	13
	3L first, *2L* second	3.971	.879	13
	total	4.100	1.118	26

Our data show a *version effect* (the *2L version* is overall faster than the *3L*), an overall *experience effect* (main effect of expertise) and an interaction between all three factors, which we label *version specific expectation effect,* and which might be related to users expectancies. These expectancies might be influenced by users exposure to different version.

4) Post-hoc Tests: To uncover the origin of the significant effects, post-hoc t-tests were computed. Differences within groups can be discovered with paired sample tests and differences between the groups with independent sample t-test. Note that the alpha level was set to 0.01 (instead of 0.05) in order to counteract alpha-error accumulation.

The paired sample test reveals the following interesting effects: a *learning through experience effect,* a *transfer effect* and a *switching effect.* When the corresponding version is presented first, for both versions, a statistical significant *learning through experience effect* is revealed by comparing the new and the expert condition. For *2L version first*, the difference between new and expert is highly significant ($t(12)=6.940$, $p<0.001$) as is the difference between new and expert for *3L version first*, with $t(12)=3.590$, $p<0.005$. The comparison between new and expert condition with the same version, but presented as second version revealed no significant effects. So for both versions performance in the third and fourth run does not improve significantly. Nevertheless, the improvement between new users and expert users is a clear indication that both versions of the application are suitable for learning.

There are significant *transfer effects* in the group *3L first, 2L second*; as there is a significant improvement between *3LN and 2LN*, with $t(12)=5.221$, $p<0.001$as well as from *3LN to 2LE* ($t(12)= 5.098$, $p<0.001$ and from *3LE to 2LE* (($t(12)=3.591$, $p<0.01$).

TABLE III. POST-HOC COMPARISON WITHIN SUBJECTS

order	conditions	t	df	sig. (2-tailed)	effectsize (dz)
2L first, *3L* second	*2LN*vs. *2LE***	6.940	12	0.000	1.924
	2LE vs. *3LN**	- 3.273	12	0.007	0.907
	2LN vs. *3LE*	1.956	12	0.074	0.542
	2LE vs. *3LE*	-1.699	12	0.115	0.471
	3LN vs. *3LE*	1.272	12	0.227	0.352
	2LN vs. *3LN*	0.795	12	0.442	0.220
3L first, *2L* second	*3LN* vs. *2LN***	*5.221*	12	0.000	1.448
	3LN vs. *2LE***	5.098	12	0.000	1.414
	3LE vs. *2LE**	3.591	12	0.004	0.995
	3LN vs. *3LE**	3.590	12	0.004	0.995
	3LE vs. *2LN*	1.537	12	0.150	0.426
	2LN vs. *2LE*	1.343	12	0.204	0.372

Note that *2L* and *3L* connote in *2L* version and *3L* version and *E* in expert and *N* in new.

For the group *2L first, 3L second* a significant *switching effect*, e.g., a drop in performance between *2LE and 3LN* is revealed (t (12) = -3.273, p<0.01).

The independent sample t-test compares conditions that do not comprise of the same users. For novice users, interacting the very first time with this application, descriptive results indicate a general advantage for the *2L version*. Nevertheless, on a statistical level, for first time users, both versions are equally difficult (*3LN1 vs. 2LN2*, t (24) =1.346, p=0.2). For expert users, without experience from a different version, the *3L version* is significantly slower, than the *2L version* (*3LE1* vs. *2LE2*, t (24) = 3.412, p<0.005).

by the data, a conceptual ACT-R model was developed that does the same task as the participants.

The model was written to search for products just as human participants do. For simplicity, only product search via the *categorical* pathway is modeled. As first step encoding of the requested product is required. Please note that this paper does not focus on visual-motor processing, and no eye tracking data is collected, we will present merely the core concept of how the models mental model changes. Nevertheless, supplementing the model with visual processes is unproblematic. This is done through goal buffer.

TABLE IV. POST-HOC COMPARISON BETWEEN SUBJECTS

conditions	t	df	sig. (2-tailed)	effectsize(ρ)
2LE2 vs. *2LN1*	-6.000**	24	0.000	2.353
2LN2 vs. *2LN1*	-5.211**	24	0.000	2.043
3LN1 vs. *2LE2*	5.181**	24	0.000	2.032
3LE1 vs. *2LE2*	3.412*	24	0.002	1.338
3LE1 vs. *2LN2*	-3.247*	24	0.003	1.273
3LN1 vs. *3LE2*	2.611	24	0.015	1.023
2LE2 vs. *3LN1*	-2.563	24	0.017	1.005
2LN2 vs. *2LE1*	2.421	24	0.023	0.949
2LN2 vs. *3LN1*	-2.021	24	0.055	0.792
3LN1 vs. *3LN2*	1.631	24	0.116	0.639
2LE2 vs. *3LE1*	-1.403	24	0.173	0.550
3LN1 vs. *2LN2*	1.346	24	0.191	0.528
3LE1 vs. *3LN2*	-1.192	24	0.245	0.467
2LN2 vs. *3LE2*	-0.819	24	0.421	0.321
2LE2 vs. *2LE1*	0.580	24	0.567	0.227
3LE1 vs. *3LE2*	-0.012	24	0.991	0.004

Note that *2L* and *3L* connote *2L* version and *3L* version. *E* means expert and *N* means new. The numbers 1 and 2 correspond to the group. Number 1 stands for the group *3L* first, *2L* second and 2 for group *2L* first, *3L* second. For example *2LE2* vs. *2LN1* is the comparison between the *2L* version expert, where the *2L* version is presented as first version (group 2) versus the *2L new* version, when the *2L* version is presented second (group 1).

But as users become more experienced with the application in general (e.g., comparing *2L expert second* with *3L expert second*) both version do not differ on a statistical level (*2LE2* vs. *3LE1*, n.s.). Table IV presents the results of a two-way t-test between the two groups. In conclusion, statistical differences between both versions are found, but for real novice users and very experiences users, both versions do not differ on a statistical level. Therefore, efficiency is the same for both versions.

5) The ACT-R Modell: In order to get a deeper understanding about the causes and mechanisms indicated

Hence, the goal buffer contains the information about what product is required. The next step represents an attempt to retrieve information from declarative memory about the *version specific category membership* of the required product.

This knowledge is represented as a chunk and consists of all vital information for finding the product in the application. Hence, *version specific category membership chunks* contain all information necessary to navigate the application. These chunks represent the mental model. The slots of the chunks contain the words leading to the product (see Fig. 3). For models new to the application, *version*

specific category membership chunks are nonexistent and for this reason the retrieval is unsuccessful. Therefore, such *version specific category membership chunks* have to be build via the general semantic knowledge. The general semantic knowledge depicts world knowledge and includes *association chunks* between products and shops and between products and subcategories. This general semantic knowledge (*association chunks*) is utilized for searching the requested product. If the retrieval of a *version specific category membership* chunk is unsuccessful, a different product search strategy is selected. Word for word, each term represented on screen, is read. For each word the attempt to retrieve an *association chunk* between the desired product and the current word is made. For example if the requested product is *alcohol free beer* and the word *bakery* is read, an attempt to retrieve a chunk that holds an association between *bakery* and *alcohol free beer* is made. If such a chunk cannot be retrieved, the next word is read, for example *beverage store* and again the attempt to retrieve a chunk that holds an association between *alcohol free beer* and *beverage store is made.* If such a chunk can be retrieved, the word is selected and a *version specific category-membership chunk* is build.

In summary, navigating the application is modeled via two types of chunks- *association chunks* and *version specific category membership chunks*. *Association chunks* represent world knowledge and contain association between different words and can be retrieved from declarative memory without prior exposure to the application. *Version specific category membership chunks* represent the mental models of the pathways leading to the products. They are built in the models imaginal buffer, during exposure with the application and can only be retrieved if the product was encountered before.

6) Discussion of the effects: Is a shallower menu-structure beneficial?

a) Empirical: The *version effect* discovered in the ANOVA indicates that the *2L version* is overall faster than the *3L version.* Post-hoc tests show no statistical difference, neither for real novice users, nor for very experienced users. On a descriptive level, as users become more familiar with the application, the benefit of a shallower version is less relevant, since the difference *2L expert* vs. *3L expert* is less than the difference between *3L new* vs. *2L new*.

P2
```
alcohol-free-beer-P2
isa product-path
first                bottleshop

final                alcohol-free-beer
```

P3
```
alcohol-free-beer-P3
isa product-path
first                bottleshop
second               beer
final                alcohol-free-beer
```

A 1
```
beer-alcohol-free-beer
isa assosiation
target               alcohol-free-beer
associate            beer
```

A 2
```
bottleshop-alcohol-free-beer
isa assosiation
target               alcohol-free-beer
associate            bottleshop
```

Figure 3. An example of different chunk-types used in the model. P2 and P3 are the version specific category membership chunks for the 2L and the 3L version, respectively. A1 and A2 are the association chunks.

The building of these chunks takes place in the imaginal buffer. Fig. 3 represents the different chunks used in the model for the product *alcohol free beer*. After the word is selected, the next page of the application opens and the process of reading and searching for *association chunks* is repeated. Furthermore, if retrievals of *association chunks* are successful, the *version specific category membership chunk* is supplemented. Finally, when the requested product is found and clicked on, all slots of the *version specific category membership chunk*s are filled with content. The chunk is then cleared from the imaginal buffer and placed into declarative memory. Thus, when the product is requested a second time, a complete product path is available and can be retrieved.

b) Explanation: A shallower menu structure requires fewer clicks than a deeper menu structure. But more interesting in this context is, the question to what extent higher level cognitive processes such as memory retrievals are responsible for the difference in product search time between the two versions. We argue that more memory retrievals can be seen as a higher amount of cognitive load. Issues concerning cognitive load, can best be answered via cognitive modeling.

c) Modeling: The building of *version specific category membership chunks* out of *association chunks* continues until all product pathways are established. This building process of *version specific category membership chunks* takes longer for *3L version* than for *2L version*. Since the *3L*

version requires more interaction steps and therefore more encoding of these steps than the *2L version*. Furthermore, for the *3L version* more retrievals of general knowledge (about which shop holds which subcategory, and which subcategory holds which product) are needed. The knowledge of subcategories is unnecessary for the *2L version*. An "expert" model can retrieve *version specific category membership chunks* for all products. Retrievals from declarative memory are much less frequently then for a "learning" model, functioning with *association chunks.* For both versions the number of retrievals is the same as soon as interaction with the application is realized solely via *version specific category membership chunks*. So, for an "experienced" model, the number of clicks alone is responsible for differences in search time.

Note: If simply motor processes (e.g., clicks) differ between the two versions, a new study should investigate if the benefit of the *2L version* is still measurable for products that require menu scrolling, or if the *3L version* may be more favorable for such products.

In general linear hierarchical applications with shallower menu structures require more scrolling than those with a deeper menu structure. It is very plausible that for expert users, a deeper menu-structure with less scrolling processes is more beneficial than a shallower menu structure, since the amount of cognitive load (memory retrievals) is the same for both versions.

a) Does Learning occur?

a. *Empirical:* The data show a clear *experience effect* as participants become more familiar with the application, the mean trial duration decreases.

b. *Modeling:* Learning, defined as the reduction of product search time, can be explained by the modeling approach as follows: As long as a mental model of the product pathway for all products is not complete, the constant retrieval of *association chunks* is necessary. For each processed word a retrieval request for an *association chunk* containing both the product and the current word is made. If such an *association chunk* cannot be retrieved the next word is read and the process is continued until an *association chunk* is found. Then the word is selected and *version specific category membership chunk* is built up. Searching, encoding and retrieving chunks take time. When *version specific category membership chunks* are available in declarative memory the number of retrievals is reduced-resulting in reduced product search time. So, a "novice" model is constantly visual searching, encoding and retrieving chunks. An "expert" model, on the other hand has the relevant knowledge about specific product pathways and therefore, less time is spent on retrieving chunks.

In conclusion, the main reason for learning is that a mental model of the application is built. This mental model consists of the relevant product pathways. As soon as the *version specific category membership chunks* can be retrieved, performance increases. Furthermore, an adequate mental model results in less cognitive load, since less memory retrievals are necessary for navigating.

a) How do version specific expectations influence performance?

Transfer effect: From the *3L* to the *2L version.*

Empirical: Performance improvements between *3L* to the *2L version* are labeled *transfer effect.*

Modeling: After repeatedly interacting with the *3L version* of the application, a mental model for the *3L version* exists. This means that *version specific category membership chunks* containing the correct product pathway for this version for all products are represented in declarative memory. The version of the application is then changed, but the task is kept the same. Without any kind of disturbance, the *3L-version specific category membership chunks* are adequate to fulfill the task with the *2L version* of the application. This is the case, since no additional information needs to be learned when switching from *3L version* to the *2L version* (note that the *3L version* includes all menu-structures of the *2L version*, but has more menu depth). Therefore, performance does not drop when switching from the *3L version* to the *2L version*.

Switching effect: From the *2L* to the *3L* version

Empirical: A *switching effect* occurs when participants familiar with the *2L version* change to the *3L version*. In the empirical data the effect is visible, in the increasing product search time from *2L first expert* to *3L second new*.

Nevertheless, participants who use the *3L version* second benefit from their experience with the *2L version* (product search time for *3L second* is lower, than for *3L first*).

TABLE V. DESIGN OF STUDY 2

order of versions	***2L* (new)**	***2L* (expert)**	***3L* (new)**	***3L* (expert)**
***2L* first, *3L* second**	Block 1	Block 2	Block 3	Block 4

Modeling: Switching from the *2L version* to the *3L version* irritates the users because they end up with a menu they did not expect and are not familiar with. In terms of the modeling approach this implicates that *2L-version specific category membership chunks* do not lead to the required product, when these are deployed with the *3L version*. On the third page of the application redemption of strategy is required and the problem is solved via *association chunks.* New *version specific category membership chunks* are built for the *3L version.* When the *3L version* is presented a second time, these chunks can be retrieved and the product search time decreases.

b) General remarks for Modeling Menu Structures: In the presented approach, mental models of product pathways for linear hierarchical menus are represented as chunks. The slot values of these chunks depict the categories, subcategories and the target in the hierarchical menu. In

order to build a mental model, *association chunks*, containing associations between words are used. These *association chunks* serve as general semantic knowledge. Performance improvements occurring when novice become expert users, are explained through the change of strategy. Novice users rely on *association chunks* and experts on chunks containing the mental model of the specific product pathway. A strategy depending on associations requires more retrievals than a mental model strategy. The more retrievals a strategy needs, the higher the cognitive load. If experienced users are confronted with a different menu hierarchy then the one they are familiar with, depending on the fashion of the new hierarchy, two opposite effects can occur. In general, if hierarchy levels are reduced, existing mental models of product pathways are still useful and productive, a *transfer effect* occurs. On the other hand, if hierarchy levels are added, new mental models of product pathways are required, a *switching effect* is found.

A. Study 2

A second study was conducted, to substantiate findings and model assumptions of the first study and also to investigate the influence of expectancies on product associations. Due to the small number of participants in the second study, statistical tests are not computed. As in the first study participants were asked to search for products with the shopping list application. The products were read to the participants and the participants had to search for the products in the application, select the products and then return to the first page. This procedure was the same, as in the first study except, that participants could only search via the stores pathway. This constraint guaranteed that participants built up a mental model of the product path from the start. The study design is another difference to the first study. In the second study all of the participants first worked with the *2L version* and then switched to the *3L version*.

1) Hypothesis

a) Learning and Version Update:

The same *experience* and *switching effects* as in the first study were assumed. We expected product search time to decrease, if the same version is used the second time and to increase after the version switches to a version with extra menu layers.

b) *The Influence of Expectations:* We predicted that longer product search times occur for category pairs that are more unfamiliar than others. We further predicted that as category affiliation become more familiar, differences in search time between products disappear.

2) Results

a) Empirical Results: As Fig. 4 shows, there is a clear switching effect (e.g., an increase in mean product searches time, after participants switch from the *2L* version to the *3L* version). Similar to the previous study, the data show a clear experience effect, with product search time decreasing as participants become more familiar with the version of the application; therefore, for both versions (*3L* and *2L*) product search in the new condition takes longer than the expert condition. To decipher how product search time differs between different products, learning curves were computed (see Fig. 5). These present the mean product search time for each individual product. With such en detail information, differences between specific products can be uncovered. The exact product search times are also presented in Table VI. In both of the new conditions strong time variations can be observed.

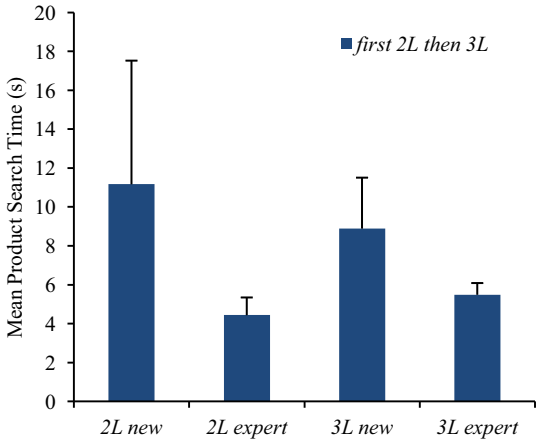


Figure 4. Mean trial time of study 2.

In the expert conditions, on the contrary only small variations between the different products exist.

A possible explanation for the observed variations in the new condition is, that some products are easier to find than others. The non presence of variations in the expert conditions indicates, that users have a correct mental model, e.g., they have learned the item labeling of the application. Hence, a qualitative explanation for product search time variations between different products will be provided.

Products that result in large search time in the new condition are the second *clabbered milk*, the third *canned pineapple* and the eighth product *gilthead*. Products with rather short product search times in the new condition are product number four *body wash* and product number seven *top-fermented dark beer*.

In post-hoc questioning the participants revealed that they expected *clabbered milk* in the *beverage store* and not in the *deli* as it was presented in the app. They also reported that they did not expect *canned pineapple* in the *corner store* and some participants were not aware that *gilthead* is a fish. A plausible explanation for variations between products is the fact, that some category pairs are more familiar for the participants than others. Higher standard deviations for the more uncommon products in the new conditions also provide evidence for this explanation (see Table II).

TABLE VI. MEAN TRIAL TIME PER ITEM FOR STUDY 2.

	new					expert			
	***3L* mean**	***3L std***	***2L* mean**	***2L std***		***3L* mean**	***3L std***	***2L* mean**	***2L std***
Alkoholfreies Bier alcohol free beer	8.918	*8.268*	10.408	*8.510*	*Alkoholfreies Bier* alcohol free beer	5.521	*1.643*	3.984	*0.907*
Dickmilch clabbered milk	10.513	*9.029*	20.596	*8.242*	*Dorade* gilthead	5.143	*1.213*	5.378	*5.758*
AnanasDose canned pineapple	10.007	*13.380*	17.091	*9.222*	*AnanasDose* canned pineapple	5.706	*2.026*	5.249	*4.337*
Duschgel body wash	7.346	*2.103*	4.844	*2.117*	*Duschgel* body wash	5.566	*1.410*	4.108	*0.747*
Amerikaner black and white cookie	11.241	*7.187*	5.585	*4.836*	*Edamer* Edam cheese	5.274	*1.242*	5.514	*2.394*
Edamer Edam cheese	7.053	*2.413*	11.114	*8.260*	*Altbier* top-fermented dark beer	4.920	*1.556*	3.895	*1.678*
Altbier top-fermented dark beer	4.717	*1.238*	3.970	*1.054*	*Acrylfarbe* acrylic paint	5.536	*1.909*	4.259	*1.706*
Dorade gilthead	13.232	*15.250*	19.189	*21.592*	*Dickmilch* clabbered milk	6.896	*2.357*	4.913	*3.161*
Acrylfarbe acrylic paint	7.024	*2.069*	7.759	*5.885*	*Amerikaner* black and white cookie	4.853	*1.034*	2.683	*1.230*

For most of the products, participants take longer with *2L version new* than with *3L version new*. This is probably due to learning transfer over the version, as discussed in the first study.

For product number four *black and white cookie* product search time with the *3L version new* is longer than with the *2L version new*. Post-hoc questioning revealed, that participants did not expect *black and white cookie* in the subcategory *danish (pastry)*.

Modeling: *The modeling approach from the first study can easily be complemented to explain the effects revealed by the learning curve. The results discussed indicate that uncommon products and product category pairs result in longer search times.*

To model such an effect, the general semantic knowledge of the model should be modulated, so that *association chunks* exists, that are correct in daily life (e.g., *clabbered milk* is associated with *beverage)* but misleading for the application. This would make it possible for the model to make errors. These errors would then result in longer product search times, if *association chunks* retrieved from declarative memory result in misleading product path. Another possibility to model longer product search times for uncommon words would be through parametric adjustments to the model. Common *association chunks* are required more often, than uncommon *association chunks*, therefore the activation of the common chunks should be higher, making unsuccessful retrievals of uncommon *association chunks* possible and therefore resulting in longer product search times.

3) Conclusion: In the second study, the *same experience effect* (experts are faster than novices) and the same *transfer effect* as in the first study are observed. These effects are even found when only *the categorical pathway* is used. This restriction ensures that the *product pathway* is represented by version specific *category membership chunks* as described in the first study.

Furthermore, an extending modeling approach offers two explanations that can account for variations in product search time in the new conditions. One explanation is that for uncommon product category pairs misleading *association chunks* are retrieved.

The other is that for unfamiliar products *association chunks* are used very rarely and therefore these chunks have a lower activation and retrieval failures occur.

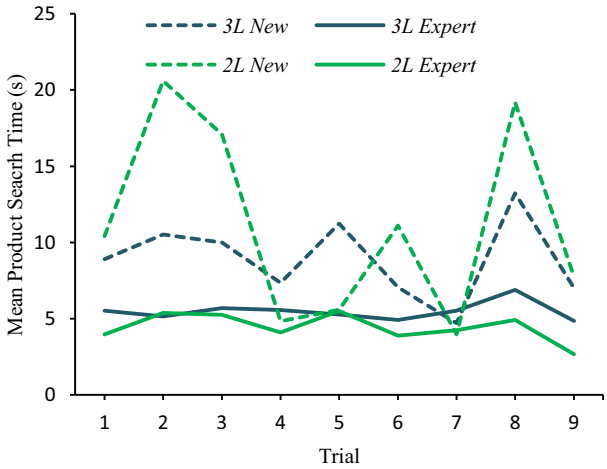


Figure 5. Mean learning curves for study 2.

Thus, designer should use only such labels for categories that are unambiguously linked to the products. Especially for first time users of an application intuitive labeling is

important. Expert users, on the other hand, can cope with uncommon labels, since they have a correct mental model.

V. DISCUSSION

A. Summary study 1

Two versions of a linear hierarchical real-life shopping list application for Android, differing in menu depth, were compared via user test. Product search time for different products gave insight into the following usability factors; efficiency, effectiveness and suitability for learning. Furthermore, novice and expert behavior and switching between both versions were investigated. The user study revealed an experience effect, namely that expert users are faster than novice users- a clear indications for suitability for learning. Overall product search with the *2L version* is faster than with the *3L* version, making the shallower version slightly more efficient. But the difference between the two versions is neither significant for users new to the application, nor for those very experienced with the application. A shallower menu hierarchy seems to be faster to handle, if the users are somewhat familiar with an application, but not jet very experienced. Both versions are effective to compose a shopping list. A *switching effect* was observed; product search time increases when switching from the *2L version* to the *3L version*. The *transfer effect*, on the other hand, is that product search time decreases when the *3L version* is presented as a first version and the *2L version* as a second version. An ACT-R based model was used to explain the results of the user study. It demonstrates how users might build up a mental model of the application. At the beginning, users need to use their general knowledge (*association chunks*) to find associations in between each processed word and the target item. Through successful navigating, they build *version specific category membership chunks* that contain the product path of the target item. These chunks represent the user´s mental model of the application. With this mental model the expert user then navigates quickly to the target item. If the different version is presented, the mental model either is still suitable for navigation (*transfer effect*) or needs to be revised (*switching effect*). Besides having an adequate mental model, the model explains the increase in performance between novices and expert, due to a reduction in cognitive load. Inexperienced users need to retrieve information about associations very frequently from their declarative memory. Each retrieval takes time. The number of retrievals is much less for experienced users, so they are faster.

B. Summary study 2

Study two was conducted to support the findings and modeling assumption of the first study; moreover, to supplement information about the nature of expectations of users new to the application. The task, application and products where kept the same, although the functionality of the application was reduced to the categorical pathway and version switches were exclusive from the *2L version* to the *3L* version. Descriptive results indicated the experience and *switching effect*. Furthermore, the evaluation of learning curves showed that in the new condition some products result in longer search times than others. An extension of the modeling approach of the first study provided two possible explanations for this: Either that the retrieval of misleading *association chunks* for uncommon product category pairs is responsible, or that too low activation of *association chunks* of unfamiliar products leads to retrieval failures.

C. Conclusion and outlook

1) Advice for designers: For a linear hierarchical menu application, that allows users to select items, the number of menu layers is not important. Especially for frequent users searching for products using a 3 or 2 layer menu, is equally efficient and effective. On the other hand, as long as users are not completely familiar with the application, a shallower menu is beneficial. It is important that these findings need to be verified with products that require scrolling. It seems plausible that more layers (resulting in more selection time) reduce the scrolling time. If version updates are necessary, designers should be aware that introducing an extra layer will reduce efficiency until users are familiar with the modification again (the switching effect). On the other hand, an update which reduces the number of layers, does not have a negative influence on efficiency (the transfer effect). Furthermore, intuitive labeling of categories is very important, especially for first time users. Designers should concentrate on finding categories, where the affiliation to the items found in these categories is immediately clear to the target population of the application. Language effects influencing potential users, such as regional terms should be considered. Besides, common categories should be preferred to uncommon ones. Though, if the scope of an application is essentially experts, with every-day exposure, intuitive labeling is less relevant. Experts can handle uncommon labels as well as common ones.

2) On the specific findings of our model: Two different chunks are used in the modeling approach- *association chunks* and version specific category membership chunks. Association chunks contain the association between the target product and categories- either shops or product categories. They represent general semantic knowledge and a novice model searches for products using its *association chunks*. Such a strategy requires much retrieval of *association chunks*. The *version specific category membership chunks* are the *mental model* of *product pathways* containing the target and the first and second category that leads to the product. Expert users can rely on these chunks to navigate through the application. The approach shows that these two different strategies are used by novice vs. expert users. The difference in efficiency observed between the *2L* and *3L* versions for learning users is explained through the different amount of *cognitive load* when the association chunk strategy is used. Using this strategy requires more retrievals for *3L* than for the *2L*

version. For experts, who rely on their mental model, both versions are equally efficient and require the same amount of *cognitive load*, since the number of retrievals of *version specific category membership chunks* is the same for both. When the other version is presented, the *version specific category membership chunks* are either still suitable for navigation (transfer effect) or need to be revised (switching effect). Variations in product search time in the new conditions are explained by the modeling approach through retrieval failures or through retrievals of non matching *association chunks*.

3) General conceptual modeling remark: This paper illustrates how usability influencing factors, such as efficiency, effectiveness and suitability for learning can be assessed with ACT-R based cognitive models. Further concepts such as users' mental model and cognitive load are evaluated, too. The mental model of the pathway of a linear hierarchical application can be modeled through chunks, which slots contain the target and the categories and subcategories. Such a mental model can be constructed via general semantic knowledge, which consists of *association chunks. Association chunks* have two slots, one for the target and one for the associated category. User expectations influence the handling of applications. We showed that unexpectedly associated word pairs can result in retrieval failures, due to low activation of these *association chunks*. This paper opted for a straight forward approach to the concept of cognitive load- the more retrievals from declarative memory was equalized with more cognitive load. Such an approach needs further validation.

4) Outlook: This paper focused on higher level cognitive processes and usability. Motor and visual processes were covered peripherally. Nevertheless, ACT-R provides possibilities to include exact visual processing of lists in its models and this should be done in the future. This is done best together with a model of the higher level mechanisms identified in this work. We provided a psychological plausible modeling approach for modeling the interaction of a smartphone application. Our approach is straight- forward, making transfer to other applications possible. In this work, the model was used to explain results obtained in user test, measuring efficiency, effectiveness and suitability for learning. The ACT-Droid tool allowed the model to directly interact with the application. In order to reach the long-term goal of using merely cognitive models to evaluate usability, other usability concepts have to be modeled. In the case of linear menu structures one should further model and investigate the following aspects: A considerable issue worthwhile to study is how differences between menu-hierarchies are affected by scrolling. When another layer was introduced to the application, we discovered a switching effect. A transfer effect occurred after removing one layer.

Besides the number of layers, other changes in the workflow of applications could be analyzed and tested for similar effects in the future, without having to rely on expansive user studies. Other important factors to include in the model are visual and motor processes. It would be helpful to evaluate these assumptions of the model with eye-tracking data. Further, it is interesting to see, at which point performance improvement of the model saturates. This would provide the opportunity to create a real expert model, for which learning behavior does not improve further. Especially a precise model of different kinds of user errors and different menus and a study focusing on mobile context is necessary for an overall model on the usability of menus. Furthermore, models of different user groups, such as elderly users, should be constructed. Such user groups would be presented through a set of parameters. For our approach to be a real alternative to user studies, it should allow for testing more complex applications. This will require implementing more actions to be simulated by the model, including scrolling, but also potentially costomization of interfaces (eg. favorits).

Cognitive load is a crucial factor for usabiliy, especially for novice users. Mobile applications are a use case for mental load evaluation. On mobile devices, with very limited space, users usually have no possibility to externalize their working memory to the device (e.g., making notes while working). Therefore, users have to rely on their working memory completely, thereby increasing the cognitive load. Complex applications are more demanding in terms of cognitive load, which is hard to measure in user studies. With cognitive modeling however, cognitive load can be assessed [35]. This provides a clear advantage of our aproach over traditional user studies.

REFERENCES

[1] N. Russwinkel and S. Prezenski, "ACT-R meets usability or why cognitive modeling is a useful tool to evaluate the usability of smartphone applications," in Proc. the Sixth International Conference on Advanced Cognitive Technologies and Applications (COGNITIVE 2014) IARIA, May 2014, pp. 62–65, ISSN: 2308-4197.

[2] J. Koetsier, "Google Play will hit a million apps in June". [Online]. Available from: http://venturebeat.com/2013/01/04/google-play-will-hit-a-million-apps-in-2013-probably-sooner-than-the-ios-app-store/ 2014.11.10.

[3] J. Nielsen, "Usability engineering". Morgan Kaufmann Pub, 1994.

[4] C. Engel, J. Herdin, and C. Maertin, "Exploiting HCI pattern collections for user interface generation" in Proc. of the 4th International Conference on Pervasive Patterns and Applications (PATTERNS 2014) IARIA, 2012, pp. 36–44, ISSN: 2308-3557.

[5] P. Kortum and S. C. Peres, "The Relationship Between System Effectiveness and Subjective Usability Scores Using The System Usability Scale," Int. J. Hum. Comput. Interact., vol. 30, no. 7, 2014, pp. 575–584, doi:10.1080/10447318.2014.904177.

[6] J. Nielsen, “User satisfaction vs. performance metrics,” 2012. [Online]. Available from: http://www.nngroup.com/articles/satisfaction-vs-performance-metrics/ 2014.11.10

[7] S. McDougall, M. Curry, and O. de Bruijn, “The effects of visual information on users’ mental models: an evaluation of Pathfinder analysis as a measure of icon usability,” Int. J. Cogn. Ergon., vol. 5, no. 2, pp. 153–178, 2001, doi: 10.1207/S15327566IJCE0501_4.

[8] D. Zhang and B. Adipat, “Challenges, methodologies, and issues in the usability testing of mobile applications,” Int. J. Hum. Comput. Interact., vol. 18, no. 3, pp. 293–308, Jul. 2005, doi: 10.1207/s15327590ijhc1803_3.

[9] R. Harrison, D. Flood, and D. Duce, “Usability of mobile applications: literature review and rationale for a new usability model,” J. Interact. Sci., vol. 1, no. 1, p. 1, 2013, doi: 10.1186/2194-0827-1-1.

[10] C. D. Wickens, “Multiple resources and mental workload,” Human Factors: J. of the Human Factors and Ergonomics Society vol. 50, no. 3, pp. 449–455, June 2008, doi: 10.1518/001872008X288394.

[11] S. Cao and Y. Liu, “Mental workload modeling in an integrated cognitive architecture,” Proc. Hum. Factors Ergon. Soc. Annu. Meet., vol. 55, no. 1, pp. 2083–2087, Sep. 2011, doi: 10.1177/1071181311551434.

[12] M. D. Byrne and R. W. Pew, “A history and primer of human performance modeling,” Rev. Hum. Factors Ergon., vol. 5, no. 1, pp. 225–263, Sep. 2009, doi: 10.1518/155723409X448071.

[13] B. E. John and S. Suzuki, “Toward Cognitive Modeling for Predicting Usability" Proc. Human-Computer Interaction, HCI International, July 2009, pp. 267-276, doi: 10.1007/978-3-642-02574-7_30.

[14] S. Möller, K.-P. Engelbrecht, and R. Schleicher, “Predicting the quality and usability of spoken dialogue services,” Speech Commun., vol. 50, no. 8–9, pp. 730–744, Aug. 2008, doi: 10.1016/j.specom.2008.03.001.

[15] J. R. Anderson, J. M. Fincham, Y. Qin, and A. Stocco, “A central circuit of the mind,” Trends Cogn. Sci., vol. 12, no. 4, pp. 136–143, Aug. 2008, doi: 10.1016/j.tics.2008.01.006.

[16] G. Bailly, A. Oulasvirta, D. P. Brumby, and A. Howes, “Model of visual search and selection time in linear menus,” Proc. 32nd Annu. ACM Conf. Hum. factors Comput. Syst. (CHI ’14), pp. 3865–3874, April 2014, doi: 10.1145/2556288.2557093.

[17] E. L. Nilsen, “Perceptual-motor control in human-computer interaction,” University of Michigan, 1996.

[18] D. E. Kieras and D. E. Meyer, “An overview of the EPIC architecture for cognition and performance with application to human-computer interaction,” Human–Computer Interact., vol. 12, no. 4, pp. 391–438, Dec. 1997, doi: 10.1207/s15327051hci1204_4.

[19] S. Lindner, P. Büttner, G. Taenzer, S. Vaupel, and N. Russwinkel, “Towards an efficient evaluation of the usability of android apps by cognitive models,” in Proc. Kognitive Systeme III, 2014.

[20] A. Cockburn, C. Gutwin, and S. Greenberg, “A predictive model of menu performance,” Proc. SIGCHI Conf. Hum. factors Comput. Syst. (CHI ’07), April 2007, pp. 627-636, 2007, doi: 10.1145/1240624.1240723.

[21] B. Mehlenbacher, T. Duffy, and J. Palmer, “Finding information on a menu: linking menu organization to the user’s goals,” Human-Computer Interact., vol. 4, no. 3, pp. 231–251, Sep. 1989, doi: 10.1207/s15327051hci0403_3.

[22] D. Ahlström, A. Cockburn, C. Gutwin, P. Irani, and I. Systems, “Why it’s quick to be square: modelling new and existing hierarchical menu designs,” Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI’10), pp. 1371–1380, April 2010, doi: 10.1145/1753326.1753534.

[23] M. Ziefle and S. Bay, “Mental models of a cellular phone menu. comparing older and younger novice users,” in Mobile Human-Computer Interaction - MobileHCI 2004 SE - 3, vol. 3160, S. Brewster and M. Dunlop, Eds. Springer Berlin Heidelberg, pp. 25–37, Sept. 2004, doi: 10.1007/978-3-540-28637-0_3.

[24] M. D. Byrne, J. R. Anderson, S. Douglass, and M. Matessa, “Eye tracking the visual search of click-down menus,” in Proc. of the SIGCHI conference on Human factors in computing systems the CHI is the limit (CHI ’99), pp. 402–409, May 1999, doi: 10.1145/302979.303118.

[25] J. E. McDonald, J. D. Stone, and L. S. Liebelt, “Searching for Items in Menus: The Effects of Organization and Type of Target,” in Proc. of the Human Factors and Ergonomics Society Annual Meeting, vol. 27, no. 9, pp. 834–837, October 1983, doi: 10.1177/154193128302700919.

[26] M. D. Byrne, “ACT-R/PM and menu selection: applying a cognitive architecture to HCI,” Int. J. Hum. Comput. Stud., vol. 55, no. 1, pp. 41–84, Jul. 2001, doi: 10.1006/ijhc.2001.0469.

[27] V. Kaptelinin, “Item Recognition In Menu Selection : The Effect Of Practice,” in CHI ’93 INTERACT '93 and CHI '93 Conference Companion on Human Factors in Computing Systems, pp. 183–184, April 1993, doi: 10.1145/259964.260196.

[28] E. Lee and J. Macgregor, “Minimizing User Search Time in Menu Retrieval Systems,” Human Factors: The J. of the Human Factors and Ergonomics Society, vol. 27, no. 2, pp. 157-162, April 1985, doi: 10.1177/001872088502700203.

[29] A. Geven, R. Sefelin, and M. Tscheligi, “Depth and Breadth away from the Desktop – the Optimal Information Hierarchy for Mobile Use,” in MobileHCI ’06 Proceedings of the 8th conference on Human-computer interaction with mobile devices and services, pp. 157–164, Sept. 2006, doi: 10.1145/1152215.1152248.

[30] K. Samp, “Designing Graphical Menus for Novices and Experts : Connecting Design Characteristics with Design Goals,” in Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13), April 2013, pp. 3159–3168, 10.1145/2470654.2466432.

[31] A. Cockburn and C. Gutwin, “A Predictive Model of Human Performance With Scrolling and Hierarchical Lists,” Human-Computer Interact., vol. 24, no. 3, pp. 273–314, Jul. 2009, doi: 10.1080/07370020902990402.

[32] A. J. Hornof and D. E. Kieras, “Cognitive modeling reveals menu search in both random and systematic,” Proc. SIGCHI Conf. Hum. factors Comput. Syst. (CHI ’97), April 1997, pp. 107–114, doi: 10.1145/258549.258621.

[33] J. R. Anderson, M. Matessa, and C. Lebiere, “ACT-R: A Theory of Higher Level Cognition and Its Relation to Visual Attention,” Hum.- Comput. Interact., vol. 12, no. 4, pp. 439–462, Sept. 1997, doi: 10.1207/s15327051hci1204_5.

[34] A. M. Collins and M. R. Quillian, “Retrieval time from semantic memory,” J. Verbal Learning Verbal Behav., vol. 8, no. 2, pp. 240–248, 1969, doi: 10.1016/S0022-5371(69)80069-1.

[35] N. Russwinkel, L. Urbas, and M. Thuering, "Predicting temporal errors in complex task environments: A computational and experimental approach," Cognitive Systems Research, vol. 12, no. 3-4, pp. 336-354, Sept. 2011, doi: 10.1016/j.cogsys.2010.09.003.

Conceptual Modeling Patterns of Business Processes

Remigijus Gustas
Department of Information Systems
Karlstad University
Karlstad, Sweden
Remigijus.Gustas@kau.se

Prima Gustiené
Department of Information Systems
Karlstad University
Karlstad, Sweden
Prima.Gustiene@kau.se

***Abstract* — System modeling patterns are similar to workflow patterns, which were established with the purpose of delineating the requirements that arise during business process modeling on a recurring basis. Traditionally, only dynamic aspects are used for the specification of modeling patterns leaving aside the static aspects of business processes. The paper presents the conceptual modeling patterns where integrity of totally different aspects can be analyzed. The advantage of such a modeling approach is that it enables visualization and integration of different modeling dimensions of information system specifications using a single diagram. Many graphical representations do not allow such visualization and integration of static and dynamic aspects. We also represent graphically interpretation of the conversation for action schema by constructs of our semantically integrated conceptual modeling method.**

Keywords-Modeling patterns; service-oriented constructs; static and dynamic aspects; sequence, iteration, synchronization, selection and enclosing patterns, universal interaction pattern.

I. INTRODUCTION

Analysis patterns are groups of concepts that represent a common construction in business modeling [7]. They are similar to workflow patterns that were originally established with the aim to define and visualize the fundamental requirements that arise during business process modeling on a recurring basis [19]. Workflow patterns are usually defined by using Business Process Modeling Notation, Unified Modeling Language (UML) Activity Diagram [16], or a Colored Petri-Net model [15]. All these notations are able to express process behavior but do not take into account the static aspects of business processes. They do not explicitly show what happens with the objects, which represent data, when some activity takes place. Integration of static and dynamic aspects is important for the control of semantic integrity among interactive, behavioral and structural aspects of a system [9]. Semantic integrity is critical to maintain the holistic representation of system specifications. To capture the holistic structure of the problem domain, it is necessary to understand how various components are interrelated. Analysis patterns presented in this paper are constructed using the principles of service orientation and they are called conceptual modeling patterns. These patterns are important for two major reasons. Firstly, they can be used for demonstration of the interplay among fundamental constructs that are used in system analysis and design process. Secondly, patterns are important for the evaluation of the expressive power of semantic modeling languages [18]. Comprehension and visual recognition of these patterns is necessary for building more specific pattern variations and composing them in different ways. Each modeling pattern language can be formally described using a set of modeling constructs and semantic rules.

Service-oriented modeling method [9] presented in this paper is based on the ontological principles [2] of the concept of service [6], and on a common understanding of the general structure of service, which is not influenced by any implementation decisions. The most fascinating idea about a service concept is that it can be applied equally well to organizational as well as technical settings. It means that the conceptual representations of service define computation independent aspects of business processes. Business processes can be seen as service compositions, which are used to specify service architecture. Service architecture can be applied for the specification of business processes in terms of organizational or technical services. Our assumption is that service-oriented representations can be communicated among business experts and system designers more effectively. Using service-oriented modeling, information systems can be structurally visualized as evolving conceptualizations of service architectures.

The concept of service in the area of information systems is mostly bound to the term of service-oriented architecture. According to Hagg and Cummings [12], Service-Oriented Architecture (SOA) is a software architectural perspective, where service is the same as component in component-based system development methodologies. SOA represents a set of guidelines and design principles, such as loose coupling, encapsulation, reuse and composability [5] [22], in which business processes can be effectively reorganized to support the business strategy [17]. From a business management perspective, SOA can provide the possibility to reach business flexibility. It enables business processes to be analyzed in terms of services. Conflicting views on the concept of service is one of the obstacles to the attempts to

develop a new science of services [3] and new academic programs focusing on services [1]. This discipline takes a broader perspective of services as opposed to technical descriptions [20].

We use the concept of service as in the sense of service science. It *"can be understood as an action or a set of actions that are performed for some value"* [21]. In the context of enterprise modeling, it is necessary to have a broader understanding and interpretation of the service concept as the definition of service goes well beyond activities that are realized using software applications. The definition of service provided by Sheth [20] emphasizes a provider - client interaction that creates and captures value. It emphasizes a value exchange between two or more parties and a transformation received by a customer [3]. The concept of service facilitates a change of business data from one valid and consistent state to another. In the public sector it sometimes denotes organizational actions. According to Ferrario and Guarino [6], services are not transferable, because they are events, not objects. The main purpose of service orientation is to capture business-relevant functionality. Taking into account the nature of the service concept, which is based on interaction between different actors to create and capture value, a service-oriented way of thinking could be applied for a computation-neutral analysis and design of business processes as well as for creation of conceptual modeling patterns.

This paper is organized as follows. In the next section, static and dynamic aspects of service interaction are described. Five different modelling modeling patterns of an integrated method are presented in the third section. In the in the fourth section, we describe a conversation for action schema and its interpretation in terms of a semantically integrated conceptual modeling method. Finally, concluding remarks are presented. This is an extended version of paper [1], which was published in BUSTECH 2014.

II. SERVICE AS AN INTERACTION

A service cannot be defined without specifying the interaction, the result of which creates value to the actors [8] involved. Service is first of all a dynamic act of doing something to somebody. It means that there are more elements necessary to construct a concept of service than just the process of 'doing'. As there are always some actors involved in such process, it signifies that it is a communication act or an interaction between human, organizational or technical components. One is asking for something and another actor provides it. The purposeful action always takes place in a service. It prescribes responsibilities for the actors involved [10]. Every business process action is goal-driven and it should always result in some value to an actor. To get the result, which provides value on demand, four key elements are necessary: service requester, service request, service performer and service response. Interrelations among these elements construct an interaction loop, which is necessary to represent service structure. Without one of these four elements, the concept of service loses its meaning. Service performers receive service requests and transform them into responses that are sent to the service requesters. Service can be characterized by an interaction loop that can be defined by a number of flows in two opposite directions. This idea is represented graphically by an elementary service interaction loop, which is delineated in Figure 1.

The main principle of service-oriented method is based on designing services as interactions among different enterprise actors. Service architecture can be represented by a composition of interaction loops. Actors in interaction loops can be seen as active elements. These elements can be organizational or technical subsystems. Organizational subsystems can be individuals, companies, divisions or roles, which denote groups of people. Technical subsystems can be represented as software or hardware components. Any coordination flow between actors [4] must be motivated by the resulting value flow. In such a way, any enterprise system can be represented and analyzed as a set of interacting loosely connected subsystems that form service architecture.

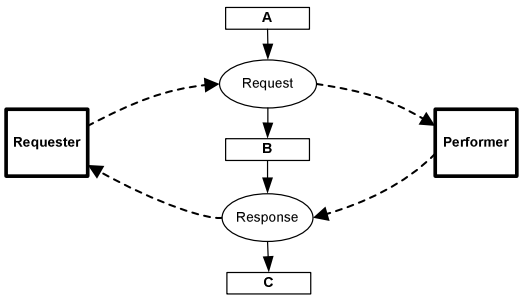


Figure 1. An elementary service interaction loop

The dynamic aspect of service includes not just interaction (▪▪▪▪▶) between actors, but also the resulting behavior among passive classes of objects when service actions are initiated. The transitions between passive classes of objects are resulting from interactions between active concepts. The internal behavior or so called objective perspective defines the dynamic aspect, which is expressed by object transitions between various classes of objects. Concepts A, B, and C define the structural aspects of data. These concepts constitute pre- and post-condition classes, which will be explained later. In such way, service modeling enables integration of business process and business data (see Figure 1).

There are two basic events for semantic modeling of service construct: creation and termination of objects [9]. These two events are used for the definition of a reclassification event, which is considered as a generic modeling construct. A creation event is denoted by an outgoing transition arrow to a post-condition class. A termination event is represented by a transition dependency directed from a pre-condition object class. Before an object is terminated, it must be created. Since a future class makes no sense for a termination event, it is not included in a

specification of action. Pre-condition class in a termination action can be understood as final during an object's life time. Reclassification of an object can be defined in terms of a communication action that is terminating an object in one class and creating it at the same time in another class. Sometimes, objects pass several classes, and then they are removed. A graphical notation of the reclassification action is presented in Figure 2.

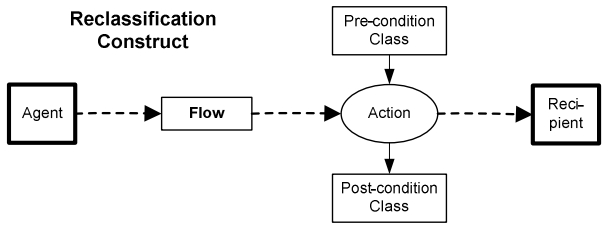


Figure 2. Graphical representation of a reclassification action

Fundamentally, three kinds of changes are possible during any transition (—►). An action is either terminating or creating an object, or it can perform termination and creation at the same time. Pre-condition and post-condition classes typically define constraints on objects, which restrict the sending and receiving of communication flows between technical or business components. A reclassification action in a computerized system can be implemented either as a sequence of one or more object creation and termination operations. Request and response flows, together with created and terminated object classes, are crucial to understand the semantic aspects of service interactions. A pre-condition object class and the input flow should be sufficient for determining a post-condition object class.

The attribute dependencies are stemming from the traditional data models. Semantics of static dependencies in object-oriented approaches are defined by multiplicities. They represent a minimum and maximum number of objects in one class that can be associated to objects in another class. We use only mandatory static dependencies from at least one side of association. A graphical notation of the attribute dependencies and their cardinalities is represented in Figure 3.

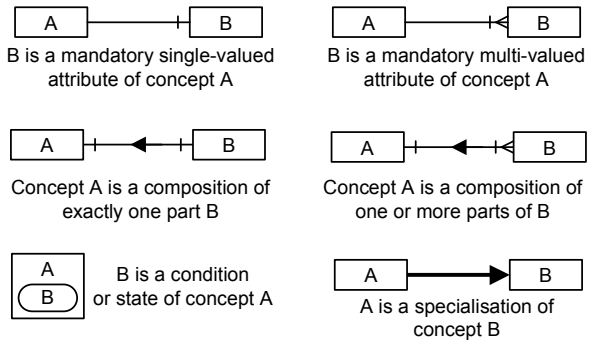


Figure 3. Graphical notation of the attribute dependencies

This notation corresponds to a classical way for representing associations between two entities [13]. One significant difference of this notation in service-oriented modeling method [9] from the traditional approaches is that the association ends are nameless. Dependencies are never used to represent association names or mappings between two sets of objects in two opposite directions. Any two concepts (in the same way as any two actors) can be linked by the attribute, inheritance or composition dependencies [9].

III. CONCEPTUAL MODELING PATTERNS

Constructs based on service orientation were used for the design of five modeling patterns. A single diagram type helps to focus on modeling integration of static and dynamic aspects. Various combinations of dependencies are able to express the main workflow control patterns such as sequence, iteration, selection, synchronization and enclosing of transaction. Ignoring the static aspects of data in the pattern modeling research creates fundamental difficulties. If just dynamic aspects are taken into consideration, then the quantity of patterns increases and their usage for business process modeling becomes more complex. Comprehensibility and visual recognition of the fundamental patterns is necessary in constructing more specific pattern variations by composing them in various ways.

Similar attributes are inherited by more specific classes according to the inheritance link (—►). Inheritance arrow denotes a specialization and generalization. Inheritance is always pointing out to a more general concept. In the diagram in Figure 4, it is possible to see two subclasses Reservation[Bill Sent] and Reservation[Paid], which are characterized by two different sets of dependencies.

We may distinguish between complete or incomplete as well as total and partial inheritance situations [24]. All these cases can be expressed by using the exclusive specialization and mutual inheritance link. Mutual inheritance dependency (◄==►) can be used for representing classes that are viewed as synonyms. It is defined as follows:
A◄==►B if and only if A—►B and B—►A.

Classification dependency (●—) specifies objects or subsystems as the instances of concepts. Classification is often referred to as instantiation, which is reverse of classification. It should be noted that classification dependency in the object-oriented approaches is a more restricted relation. It can be only defined between an object and a class. A class cannot play a role of meta-object, which is instantiated in another class.

Any class A can be viewed as an exclusive generalization of concepts B and C. A concept can be specialized by using a notion of state. For instance, *Payment* is specialized by *Payment*[*Confirmed*] as a result of confirm payment action. Various states of *Reservation* concept such as *Bill Sent* and *Paid* are also represented in Figure 4.

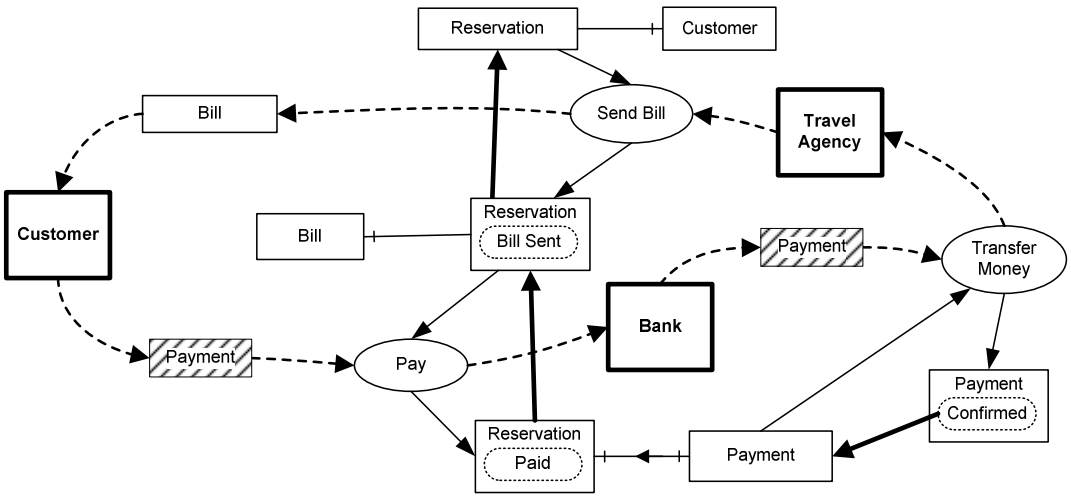


Figure 4. Example of static and dynamic dependencies

This example shows an interaction flows among three actors: Travel Agency, Customer and Bank. Interaction dependency (A ▪▪▪▶B) indicates that one actor (A) depends on another actor (B). Rectangles with shaded background are used for denotation of resource flows and light rectangles indicate information flows. Three communication actions *Send Bill*, *Pay*, and *Transfer Money*, which are triggered in a sequence, are used to express the business process of payment. *Pay* action can be executed only if the *Send Bill* action has been completed. It uses a *Reservation*[*Bill Sent*] object and produces a *Reservation*[*Paid*] object. When *Customer* receives the bill, he initiates the *Pay* action, which creates a new object *Reservation*[*Paid*] and links it to the specific *Payment*. A *Reservation*[*Paid*] consists of the compositional object of *Payment*. *Transfer Money* action can be executed only if the process of payment has been confirmed. So, according to this example, every action creates new object links that are associated with the post-condition object class. Since the post-condition class of Payment Confirmed is linked with the pre-condition class Payment by the inheritance link, the initial object is not terminated. Removal of objects in more general classes with their own attributes should occur if they are not preserved by the created objects. For instance, the missing inheritance arrow from the post-condition class would justify termination of a post-condition class object. Note that a Reservation object is required to be created in advance by another service, which is not presented in this example.

Service architecture can be composed of various interaction loops. The semantics of such composition is defined by using two or more constructs of the basic action (Figure 2). The composition of these three types of constructs can be used for the conceptualization of a continuous or finite lifecycle for one or more objects in the service interaction loop. A lifecycle of an object is typically represented by an initial, intermediate and final class. A creation event corresponds to a starting point and removal action – to the end point in an object's lifecycle. The most critical issue in the modeling of the interaction details is the semantic integrity of static and dynamic aspects. It is not sufficient to represent what type of objects are created and terminated. Service-oriented models must clearly represent attributes that must be either removed or preserved in any creation, termination and reclassification action. This is crucial to ensure the consistency of integrity constraints.

A. Sequence pattern

A pattern of sequence is a special case of an elementary interaction loop, which was presented in Figure 1. It consists of a request and response. A service request creates an object of type B, which in the second communication action is reclassified to the object of type C. These two actions are performed in a sequence and are represented in Figure 5.

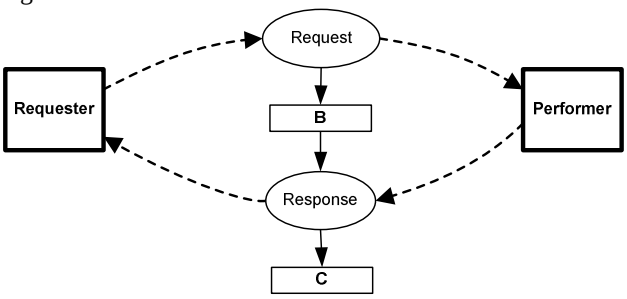


Figure 5. Sequence pattern

This pattern is used for representation the succession of events. For example, customer may order the goods by

creating a purchase order. If the goods are available, a vendor accepts the purchase order. The example of sequence pattern is represented in Figure 6.

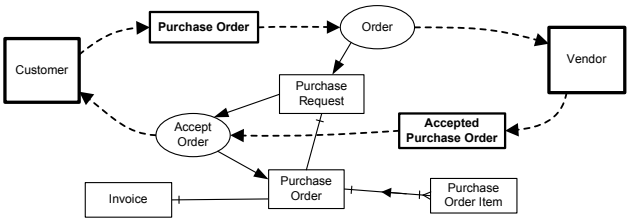


Figure 6. Example of a sequence pattern

A newly created purchase order is defined by three properties: Invoice, a set of Purchase Order Items and Purchase Request, which are necessary for delivering the order. It is not specified what will happen after that. Either the customer may withdraw the order, or it must be delivered.

B. Iteration pattern

Iteration pattern is a special case of a sequence pattern. It consists of one creation action and one removal action. The first action creates an object, which is subsequently removed. The iteration pattern is represented in Figure 7.

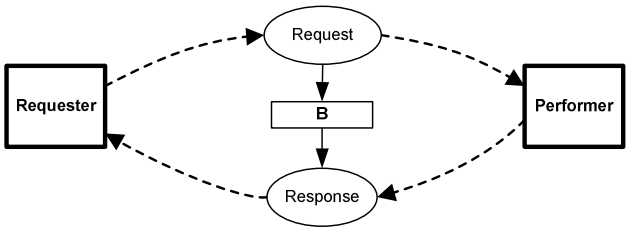


Figure 7. Iteration pattern

This pattern can be used for the representation of events that are repeated a number of times. For example, customer may order goods, which are not available. In this case, the vendor rejects the purchase order by removing it from existence. The message about the rejected purchase order is sent to the customer. The example of iteration pattern is represented in Figure 8.

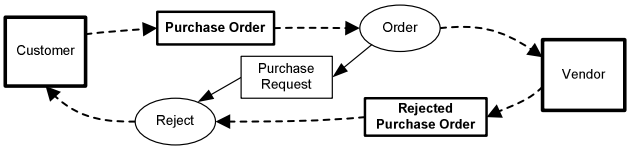


Figure 8. Example of iteration pattern

As we can see, when the Purchase Request is removed, the Customer may initiate a new the Purchase Order again. This interaction loop can be repeated a number of times. The diagrams that are represented in Figure 6 and Figure 8 can be superimposed into the single diagram. In this way, we can see what kind of alternative actions are available to the actors involved.

C. Synchronization pattern

A synchronization pattern is used when some activities must be performed concurrently. This pattern combines two parallel paths of activities. Both paths must be completed before the next process can take place. The primary interaction loop is composed of a more specific loop on a lower level of granularity. In this case, a service interaction loop on the lower layer of decomposition is viewed as an underlying interaction loop. The execution of the underlying loop must be synchronized with the primary interaction loop. The synchronization pattern is presented in Figure 9.

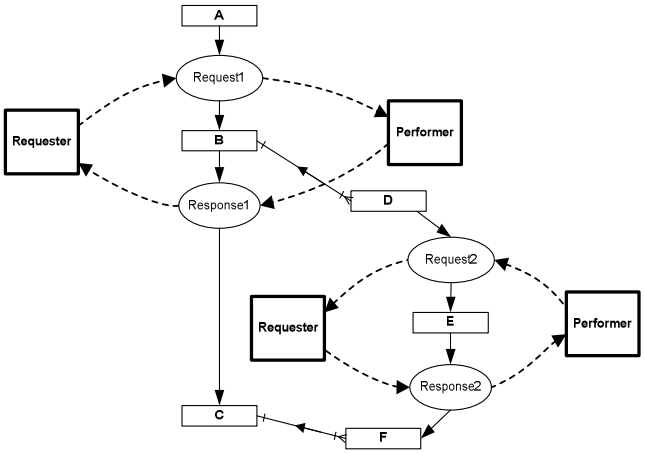


Figure 9. Synchronization pattern

This pattern illustrates that the action of Request1 creates a compositional object B, which consist of parts D. At least one part D must be created. Then object B is reclassified to C, object D must be also reclassified to E and then to F. If a compositional object is created, then the parts are created as well. If a compositional object is removed, then the parts are terminated at the same time. That is the reason why the action is propagated from a whole to a part according to the rule of class composition. The propagation of actions is a useful modeling quality. It allows a natural modeling of concurrency. Synchronization pattern is similar to concurrent activities (fork and merge of control) in an activity diagram [16].

The graphical example of synchronization is illustrated in Figure 10. In this example, the object reclassification effects represent the important semantic details of an unambiguous scenario in which three interaction loops are combined. Create Reservation action propagates to parts on the lower level of abstraction. Termination of *Hotel Reservation Request* requires termination of *Hotel Room*[*Desirable*]. Creation of *Hotel Reservation* requires

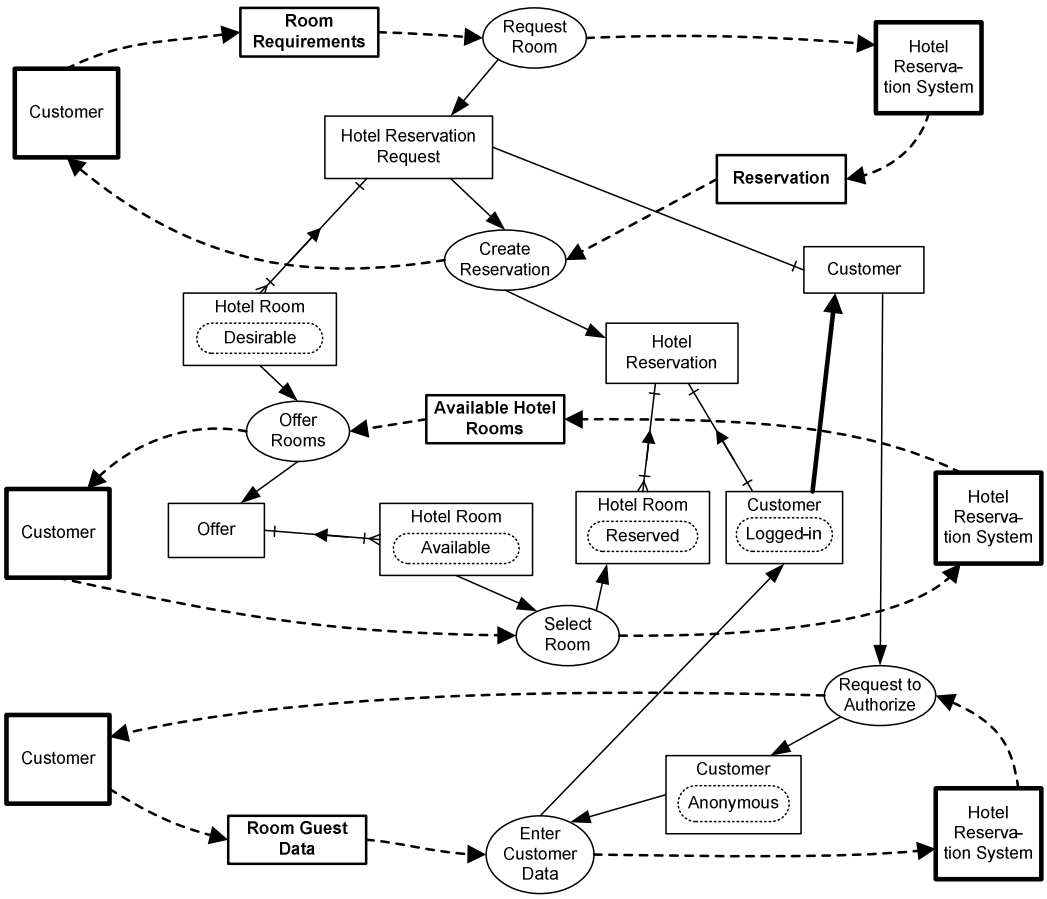


Figure 10. Example of a synchronization pattern

creation of one or more *Hotel Room*[*Reserved*]. According to the presented diagram, the underlying interaction loop action *Select Room* can be reiterated more than once, because *Hotel Reservation* is defined as the composition of one or more *Hotel Room*[*Reserved*].

The underlying interaction loop describes a Customer's response to the Hotel Reservation System's request. If a customer expects to receive a Reservation flow from the Hotel Reservation System, it is necessary for him to get a reply in the underlying loop from the technical component. The request and reply of the second underlying loop is specified as follows:
If Offer Rooms (Hotel Reservation System ▪▪▪▶ Customer), then Select Room(Customer ▪▪▪▶ Hotel Reservation System).

The actions of the underlying loop are synchronized with the primary interaction loop. According to the presented description, Create Reservation is a reclassification action, which is composed of the Offer Rooms and Select Rooms actions on the lower granularity level. The Select Room action cannot be triggered prior to the Offer Rooms action. It can be performed several times for each *Hotel Room*[*Available*]. *Hotel Reservation* is a compositional object. When it is created, such parts as *Hotel Room*[*Reserved*] and *Customer*[*Logged-in*] must be created as well. The first underlying loop is necessary for offering available rooms and selecting of a desirable room. Creation of *Customer*[*Logged-in*] object requires to initiate Request to Authorize and Enter Customer Data actions that are represented by the second underlying loop.

This modeling pattern is similar to a synchronization, which can be defined by fork and merge of control in UML activity diagram.

D. Selection pattern

The *Selection* pattern can be expressed using a composition of two different sequences between the same two actors. It represents two alternative outcomes of a service request that can be selected by service performer. Two possible ways of replying by performer are mutually exclusive. Only one type of response is expected by a service requester. If the first alternative is rejected, then the performer is trying to invoke the second alternative. The selection pattern was previously published and it can be found in [11]. It is similar to branches in UML [16]. The selection pattern is represented graphically in Figure 11.

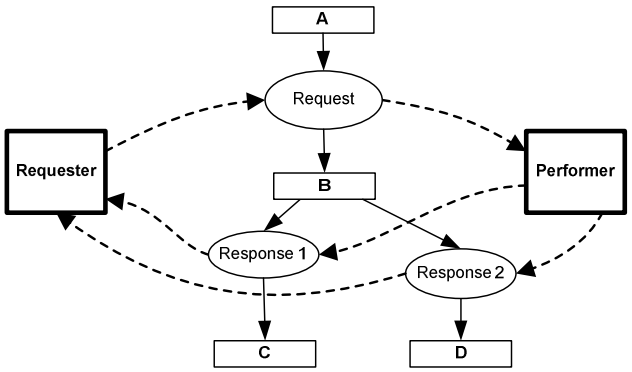


Figure 11. Selection pattern

Response 1 and Response 2 are two exclusive actions of a service performer. If Response 1 is initiated, then a pre-condition class object B is removed and a post-condition class C is created. If Response 1 has failed, then Response 2 is triggered, which reclassifies object B to D. The example of selection pattern is represented in Figure 12.

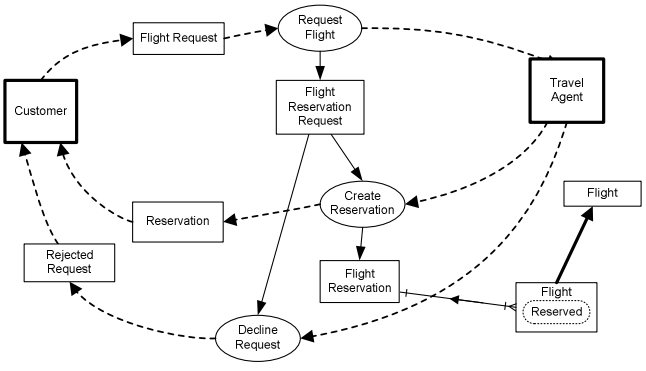


Figure 12. Example of a selection pattern

The selection pattern in the presented example can be explained as follows. The *Flight Reservation Request* is created and then it is reclassified into *Flight Reservation* in the *Create Reservation* action from the *Travel Agent*. If *Travel Agent* cannot create a *Flight Reservation*, then the alternative action of Decline Request is taking place. In this case, the *Flight Reservation Request* is terminated and a flow of Rejected Request is sent to the *Customer*. This action allows the *Customer* to reiterate the search again.

This pattern reminds us alternatives in UML, which are typically described as branches in activity diagram.

E. Enclosing pattern

An *enclosing pattern* is defined by a primary and a secondary interaction loop between requester and performer. In carrying out the work, a performer may play the role of requester in the secondary interaction loop by initiating further interactions. In this way, a network of loosely coupled actors with various roles comes into interplay to fulfill the original service request. Organizational systems may be composed of several interaction loops, which are delegated to more specific components. Enclosing pattern is similar to the enclosing of a transaction [4]. An enclosing pattern is represented graphically in Figure 13.

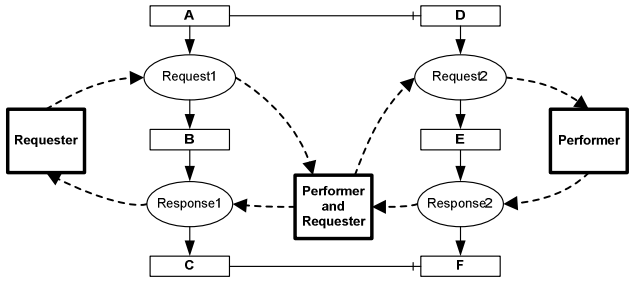


Figure 13. Enclosing pattern

The primary interaction loop consists of Request1 and Response1 actions. For the creation of object B in the primary loop, it is necessary to create its property E in the secondary loop. The reclassification of object B to C requires the removal of E and creation of F. So, the enclosing loop cannot be completed if the secondary loop is not finalized. The example of the enclosing pattern is represented in Figure 14.

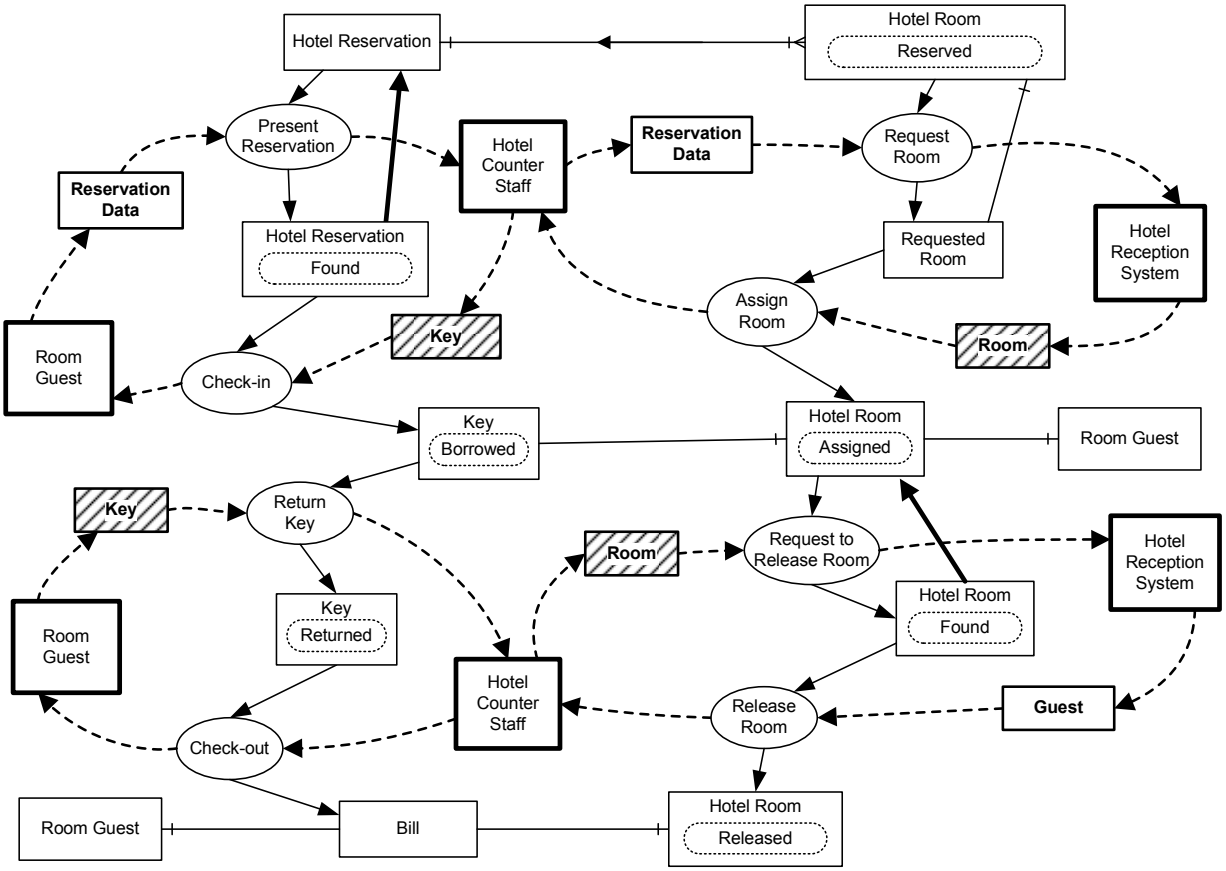


Figure 14. Two examples of enclosing pattern

If a *Room Guest* wants to *Check-in*, he needs to *Present Reservation* to the *Hotel Counter Staff*. If the hotel room is ready, the *Hotel Reception System* assigns this room to the hotel guest and produces the key, which is given to *Hotel Counter Staff*. The *Assign Room* action is executed by the *Hotel Reception System*, which is playing the role of software component. The *Check-in* action is performed by *Hotel Counter Staff*, which is playing the role of the organizational component. There is one enclosing and one enclosed interaction loop, which is represented in Figure 14. The primary interaction loop between Room Guest and Hotel Counter Staff encloses the secondary interaction loop between the Hotel Counter Staff and the Hotel Reception System. So, the *Assign Room* action is considered as a part of the *Check-in* action.

The business process, which is represented in Figure 14, consists of four interaction loops. The first primary loop is an organizational process. The secondary loop corresponds to a computerized process, which creates a *Hotel Room*[*Assigned*] object, and connects it with the *Room Guest* and *Key*[*Borrowed*] objects. The second primary loop is necessary for returning a key and checking-out a guest. It corresponds to an organizational process. The enclosed loop, which is initiated by a *Hotel Counter Staff*, corresponds to a computerized process. It is necessary for finding and releasing an assigned room.

IV. THE EXTENDED UNIVERSAL PATTERN OF INTERACTION

Interaction dependencies are extensively used in the context of enterprise engineering methods [4]. These methods are rooted in the interaction pattern analysis and the philosophy of language. The underlying idea of interaction pattern analysis can be explained by a well-known conversation for action schema [23]. The purpose for introducing this schema was initially motivated by the idea of creating computer-based tools for conducting conversations. Our intention is to apply the interaction dependencies as they are defined by the semantically integrated conceptual modeling approach [9] in combination with conventional semantic relations, which are used in the area of system analysis and design. Interaction loops can be expressed by the interplay of coordination or production events, which appear to occur in a particular pattern. This pattern is represented in Figure 15.

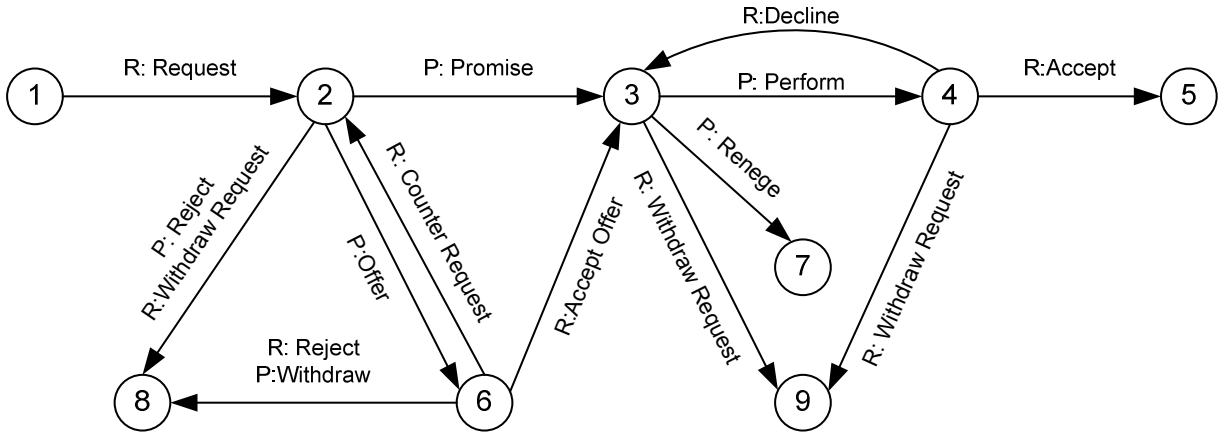


Figure 15. Conversation for action schema (Winograd & Flores, 1986)

The idea behind a conversation for action schema can be explained as turn-taking. Any service interaction pattern can be characterized by the same four types of main events, which compose a basic transaction pattern:

a) Request,
b) Promise,
c) Perform and
d) Accept.

The Requester (R) initiates a request (R:Request) action and then waits for a particular promise (P:Promise). Request, promise and acceptance are typical coordination actions, which are triggered by the corresponding types of basic events. Coordination events are always related to some specific production event, which is represented by P:Perform. Both coordination and production events can be combined together into scenarios, which represent an expected sequence of interactions between requester and performer. We will show how creation, termination or reclassification constructs of the semantically integrated conceptual modeling method can be used to define the new facts, which result from the main types of events of the basic transaction pattern.

Various interaction alternatives between two actors can also be defined by interaction dependencies, which may produce different, similar or equivalent behavioural effects. A provider may experience difficulties in satisfying a request. Instead of promising, the service provider may respond by rejecting the request. For example, the hotel reservation system may reject a request of a customer, because it is simply incorrect or incomplete. The Requester may also express disappointment in the result and decline it. Decline is represented by the termination of Result and the creation of a Declined Result object. For instance, the hotel guest may decline the assigned hotel room, which was assigned by the provide hotel room action. In this case, the basic transaction pattern can be complemented by two dissent patterns. This extended schema is known as the standard pattern [4].

In practice, it is also common that either requester or performer is willing to completely revoke some events. For example, a requester may withdraw his own request, a performer may withdraw his promise, a performer may cancel his own stated result or a requester may cancel his own acceptance. These four cancellation patterns may lead to partial or complete rollback of a transaction. These four options, which are known as cancellation patterns, should be integrated into a universal interaction pattern. A provider may also create new Offer on a basis of created Request, which can transformed into a counter request or it can be accepted by requester. All these possible outcomes are represented in the extended universal pattern, which is shown in Figure 16.

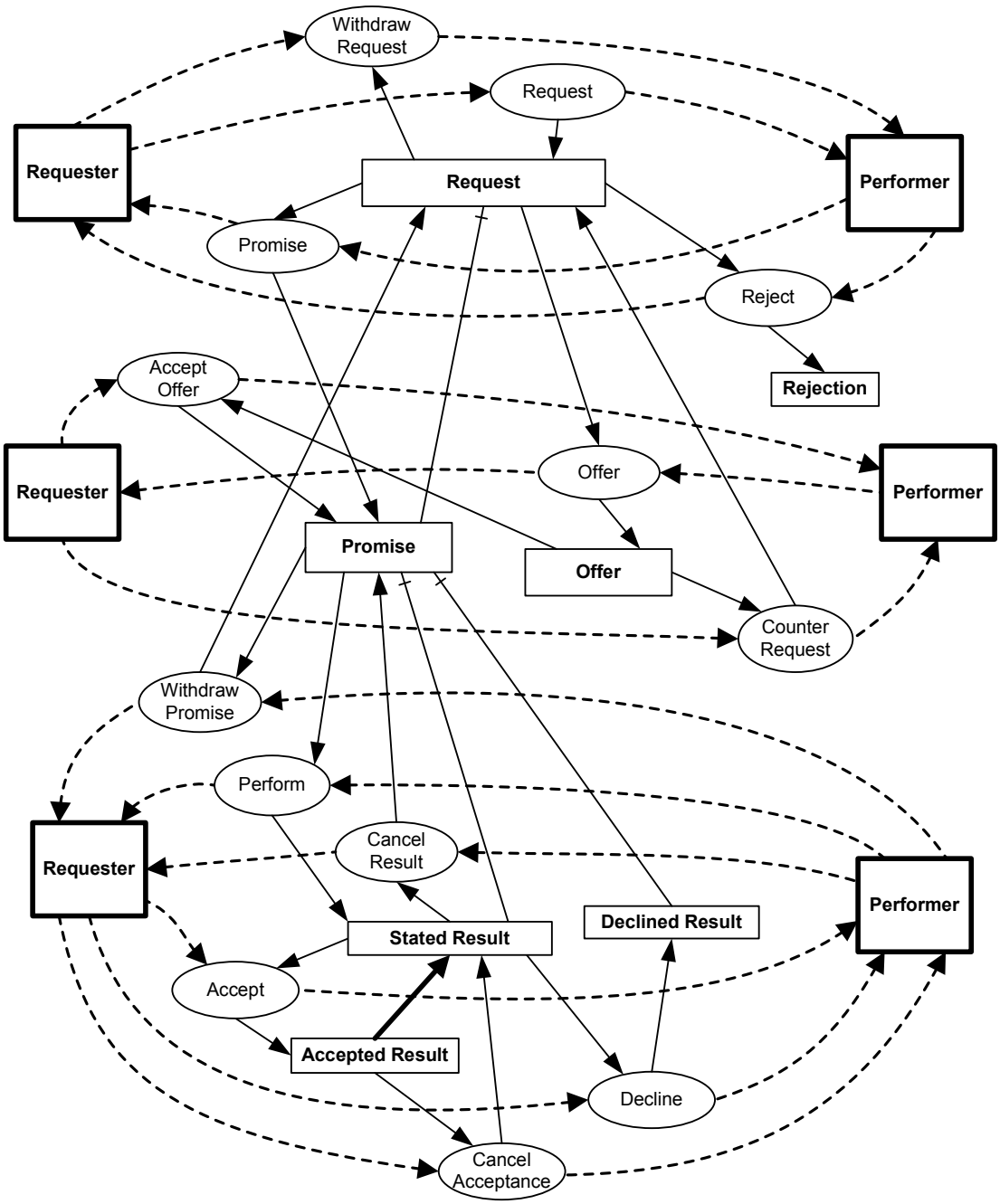


Figure 16. The extended universal interaction pattern

The presented diagram includes the standard transaction pattern and four cancellation patterns, which were analysed by Dietz [4]. It also includes an offer and counter request actions, which are taken from the conversation for action schema [23]. Every cancellation action can be performed if the corresponding fact exists. For instance, the Withdraw Request action can be triggered, if a request object was created by the Request action. Request cancellation event may occur when the customer finds a better or cheaper room in another hotel. A Withdraw Promise action may take place if a Promise for some reason cannot be fulfilled by Performer. For instance, a Hotel Room was damaged as a consequence of some unexpected event. The requester may agree or disagree to accept the consequences of the Withdraw Promise action. Please note that Withdraw Promise action terminates the Promise object and preserves

the Request object. So, the Requester will be forced to cope with four possible alternatives of communication actions such as Promise, Reject, Offer or Withdraw Request. These four alternatives are clearly visible in our new universal interaction pattern.

The third cancellation event is represented by the option Cancel Result. It can be initiated by Performer to avoid the Decline action by requester. The requester typically allows cancelling the result, because after this action the Promise is not terminated. The forth cancellation event may take place when the whole transaction was completed, but the service requester discovers some hidden problem and he regrets acceptance. For instance, the customer may try to Cancel Acceptance of the Hotel Room for the reason that wireless Internet access fails to work properly. The possibility to superimpose four cancellation patters on the standard pattern is not the only advantage of the presented modelling approach. It has sufficient expressive power to cover other special cases, which do not match the universal pattern [4].

V. CONCLUDING REMARKS

The goal of this paper was to demonstrate how the suggested service-oriented constructs can be used for the creation of five different modeling patterns. Traditionally, modeling patterns are constructed taking into account just dynamic aspects of business processes. The advantage of the suggested modeling constructs is that they allow integration of both static and dynamic aspects. One of the main contributions of this paper is the presentation of the extended universal interaction pattern.

The separation of static and dynamic details of the presented patterns creates fundamental difficulties for two major reasons:

1) Since the static aspects must somehow be compensated by using dynamic constructs, the number of patterns becomes bigger than is really necessary. Sometimes, the pattern differences are difficult to understand and they are visually unrecognizable by business experts.
2) If static aspects are not taken into account, then patterns will become more complicated to use them for the purpose of blending enterprise and software engineering.

Interaction dependencies, which define the interplay of coordination or production events, are lying in the foreground of the presented semantically integrated conceptual modeling method. It was demonstrated how interaction dependencies can be analyzed in interplay with the traditional semantic relations in the area of system analysis and design. However, a more systematic comparison with the well-established conceptual modeling languages is necessary. In our future work, we also intend to apply and to validate the method by more realistic trials in industry. The communication for action schema and the extended universal interaction pattern are not fully integrated. So, we need to do more research, which leads to complete integration of these two schemes.

The semantics of service architecture can be defined by using one or more interaction loops. Each interaction loop is composed of creation, termination or reclassification actions. By matching the interaction dependencies from requesters to providers, one can explore opportunities that are available to different actors. The static dependencies define complementary semantic details, which are important for reasoning about service interactions. The examples of corresponding behavior are presented in this paper as well. The novelty of such a way of modeling is that it enables integration of static and dynamic aspects, which are important to maintain a holistic representation of information system specifications. Service-oriented way of modeling is computation-neutral. Diagrams follow the basic conceptualization principle in representing only computationally neutral aspects that are not influenced by any implementation solutions. Since computation-neutral representations are easier to comprehend for business experts as well as system designers, they facilitate understanding and can be used for bridging a communication gap among different types of stakeholders.

REFERENCES

[1] R. Gustas and P. Gustiene, “Three Conceptual Modeling Patterns of Semantically Integrated Method,” *The 4-th International Conference on Business Intelligence and Technology, BUSTECH 2014*, IARIA, pp 19-24.

[2] M. A. Bunge, *Treatise on Basic Philosophy, vol.4, Ontology II: A World of Systems*, Reidel Publishing Company, Dordrecht, Netherlands, 1979.

[3] H. Chesbrough and J. Spohrer, "A Research Manifesto for Services Science," *Communications and ACM, 49*(7), 2006, pp. 35-40.

[4] J. Dietz, Enterprise Ontology: Theory and Methodology, Springer, Berlin, 2006.

[5] T. Erl, Service -Oriented Architecture: Concepts, Technology, and Design. New Jersey: Pearson, 2005.

[6] R. Ferrario and N. Guarino, "Towards an Ontological Foundation for Service Science," *Future Internet-FIS2008:The First Internet Symposium, FIS 2008 Vienna, Austria. Revised Selected Papers*, Berlin: Springer, 2008, pp. 152-169.

[7] M. Fowler, *Analysis Patterns: Reusable Object Models*. Menlo Park: Addison-Westley, 1997.

[8] J. Gordijn, E. Yu, and B. van der Raadt, "e-Service Design Using i* and e3 value Modeling," *IEEE Software, 23*(3) 2006, pp. 26-33.

[9] R. Gustas and P. Gustiene, "Conceptual Modeling Method for Separation of Concerns and Integration of Structure and Behavior," *International Journal of Information System Modeling and Design*, vol. 3 (1), New York: IGI Global, 2012, pp. 48-77.

[10] S. Alter, "Service System Fundamentals: Work System, Value Chain, and Life Cycle," *IBM Systems Journal, 47*(1), 2008, pp. 71-85.

[11] P. Gustiené, Development of a New Service-Oriented Modeling Method for Information Systems Analysis and Design, PhD Thesis, Karlstad University Studies, 2010:19, 2010.

[12] S. Hagg and M. Cummings, Managing Information Systems for the Information Age. New York: McGraw-Hill, 2008.

[13] J. A. Hoffer, J. F. George and J.S. Valacich, *Modern Systems Analysis and Design.* New Jersey: Pearson, 2004.

[14] I. Jacobson and P. W. Ng, *Aspect-Oriented Software Development with Use Cases*. New Jersey: Pearson, 2005.

[15] K. Jensen, Coloured Perti Nets. Basic Concepts, Analysis Methods and Practical Use. *Monographs in Theoretical Computer Science, 1, 1997*.

[16] OMG, *Unified Modeling Language Superstructure, version 2.2.* Retrieved March 7, 2014, from www.omg.org/spec/UML/2.2/.

[17] M. P. Papazoglou and W. J. van den Heuvel, "Service-Oriented Design and Development Methodology," *Journal of Web Engineering and Technology, 2*(4), 2006, pp. 412-442.

[18] A. A. Rad, M. Benyoucef and C. E. Kuziemsky, "An Evaluation Framework for Business Process Modelling Languages in Healthcare," *Journal of Theoretical and Applied Electronic Commerce Research, 4*(2), 2009, pp. 1-19.

[19] N. Russell, A. H. M. Hofstede, W. M. P. Aalst and N. Mulyar, "Workflow Control-Flow Patterns: A Revised View," (BPM Centre Report BPR-06-22). Retrieved March 5, 2014 from www.workflowpatterns.com/documentation/documents/BPM-06-22.pdf.

[20] A. Sheth, K. Verma and K. Gomadam, "Semantics to Energize the Full Service Spectrum," *Communications of the ACM, 49*(7), 2006, pp. 55-61.

[21] P. Spohrer, P. Maglio, J. Bailey and D. Gruhl, "Steps Towards a Science of Service Systems," *IEEE Computer 40*(1), 2007, pp. 71-77.

[22] O. Zimmerman, P. Krogdahl and C. Gee, "Elements of Service-Oriented Analysis and Design," Retrieved March 6, 2014 from wsdl2code.googlecode.com/svn/trunk/06-CD/02-Literatur/Zimmermann%20et%20al.%202004.pdf.

[23] T. Winograd and R. Flores, Understanding Computers and Cognition: A New Foundation for Design, Ablex Norwood, NJ, 1986.

[24] M. Blaha and J. Rumbaugh, Object-Oriented Modeling and Design with UML, Pearson Prentice Hall, 2005.

Assessing the Difficulty of Chess Tactical Problems

Dayana Hristova*, Matej Guid, and Ivan Bratko

Faculty of Computer and Information Science
University of Ljubljana
Ljubljana, Slovenia
* On leave of absence from University of Vienna, Vienna, Austria

***Abstract*—We investigate experts' ability to assess the difficulty of a mental task for a human. The final aim is to find formalized measures of difficulty that could be used in automated assessment of the difficulty of a task. In experiments with tactical chess problems, the experts' estimations of difficulty are compared to the statistic-based difficulty ratings on the Chess Tempo website. In an eye tracking experiment, the subjects' solutions to chess problems and the moves that they considered are analyzed. Performance data (time and accuracy) are used as indicators of subjectively perceived difficulty. We also aim to identify the attributes of tactical positions that affect the difficulty of the problem. Understanding the connection between players' estimation of difficulty and the properties of the search trees of variations considered is essential, but not sufficient, for modeling the difficulty of tactical problems. Our findings include that (a) assessing difficulty is also very difficult for human experts, and (b) algorithms designed to estimate difficulty should interpret the complexity of a game tree in the light of knowledge-based patterns that human players are able to detect in a chess problem.**

Keywords–Task Difficulty; Assessing Problem Difficulty; Eye Tracking; Problem Solving; Chess Tactical Problems; Chess

I. INTRODUCTION

In this article, we investigate the ability of experts to assess the difficulty of a mental task for a human, and study the possibilities for designing an algorithmic approach to predicting how difficult the problem will be to solve by humans [1], [2]. Modeling the difficulty of problems is a topic becoming increasingly salient in the context of the development of intelligent tutoring systems [3], neuroscience research on perceptual learning [4], and dynamic difficulty adjustment (DDA) for gaming [5], [6]. However, as-of-yet there is no developed methodology to reliably predict the difficulty for a person of solving a problem. This work therefore seeks to explore different ways of assessing difficulty, including human experts, and statistical analysis of performance data.

In our study, we use chess as an experimental domain. In our case, a problem is always defined as: given a chess position that is won by one of the two sides (White or Black), find the winning move, or *a* winning move in cases when several moves lead to victory. A chess problem is said to be *tactical* if the solution is reached mainly by calculating possible variations in the given position, rather than by long term positional judgement with little calculation of concrete variations. The starting point of our investigation is scrutinizing the relationship between a player's chess expertise and their ability to assess the difficulty of a tactical problem.

The term 'difficulty' requires further explanation. We are primarily concerned with *task difficulty*, which mediates between "subjective experience of difficulty" (that cannot be objectified) and "task complexity" (an inherent quality of a task; e.g., the properties of its state space). We define the difficulty of a problem as the probability of a person failing to solve the problem. Solving a problem is associated with uncertainty. Even in the case that a person solving a problem has complete knowledge relevant to the problem, she may occasionally miss the solution. In chess, there are well known cases of blunders when a chess grandmaster failed to see an obvious winning move. Accordingly, the difficulty depends on both the problem and the person.

The more experienced the person is in the area of the problem, the easier the problem is for that particular person. For a group of people of similar expertise and problem-solving skills, the problem's difficulty will be similar for all of them. In such cases, when talking about difficulty, we may leave out the reference to any particular individual within the group. We thus make the following assumption regarding the ranking of problems according to difficulty. For two people with different experience in the problem area, the ordering of two problems according to difficulty is the same for both people. That is, if problem 1 is easier than problem 2 for person A, then problem 1 is also easier than problem 2 for person B. Of course, this assumption may be debated, but we believe it is true in large majority of cases.

The aim of our investigation is to find underlying principles of difficulty perception and estimation for a defined group. This will allow us to omit the reference to individual persons and to focus on regularities that are required for modeling the difficulty of particular tasks.

In the case of chess tactical problems, human players will encounter difficulty when the problem exceeds the limitations of their cognitive abilities, i.e., their ability to detect relevant motifs and to calculate variations in [7]. The perception of difficulty can also be influenced by psychological factors, and from the way a particular problem is presented [8]. De Groot [9] and Jongman's [10] are among the first contributions to the academic research on thinking processes in chess. Both authors focus on the ability of players of different expertise to memorize chess positions. Research on expertise in chess has been mostly focused on the perceptual advantages of experts over novices [11], [12], [13], [14], [15].

Our study aims to explore the connection between task difficulty and expertise, as well as the variability among individuals. Although relatively little research has been devoted to the issue of problem difficulty, it has been addressed within the context of several domains, including Tower of Hanoi [16], Chinese rings [17], 15-puzzle [18], Traveling Salesperson Problem [19], Sokoban puzzle [20], Sudoku [21], and also

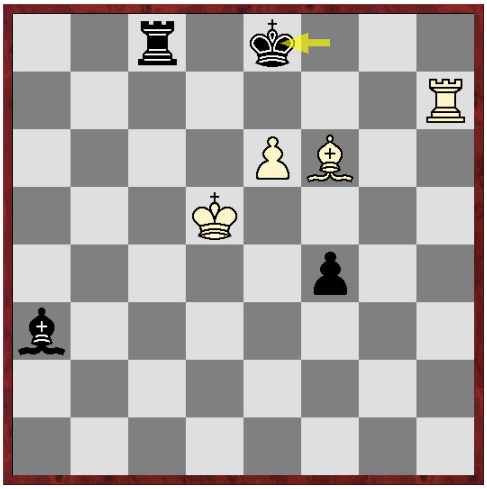


Figure 1. Chess Tempo: White to move wins. Black's last move: Kf8-e8.

chess [2]. To the best of our knowledge, no related work has been focused either on domain experts' abilities to estimate the difficulty of a mental task for a human, or on modeling the difficulty of chess tactical problems.

To approach task difficulty we are using performance measures (accuracy of solution, time, variations considered, ranking positions), psychophysiological measures (eye tracking), and qualitative retrospective reports (on perceived difficulty and on variations considered). The paper is organized as follows. In Section II, we introduce the difficulty ratings, state our hypothesis and explain why modeling the difficulty of chess tactical positions is problematic. Section III describes our methodology. We present our results of experimental data analysis in Section IV, which is followed by a thorough discussion of illustrative examples from the eye-tracking experiment. The final section of the paper is reserved for concluding remarks and directions for future work.

II. Toward Modeling Difficulty

A. Difficulty Ratings

We have adopted the difficulty ratings of Chess Tempo – an online chess platform available at www.chesstempo.com – as a reference. The Chess Tempo rating system for chess tactical problems is based on the Glicko Rating System [22]. Problems and users are both given ratings, and the user and problem rating are updated in a manner similar to the updates made after two chess players have played a game against each other, as in the Elo rating system [23]. If the user solves a problem correctly, the problem rating goes down, and the user's rating goes up, and vice versa: the problem's rating goes up in the case of incorrect solution. The Chess Tempo ratings of chess problems provides a base from which to analyze the ability of human experts to estimate the difficulty of a problem, and in our case to predict the statistically calculated measure of difficulty.

Fig. 1 shows an example of a Chess Tempo tactical problem. Superficially it may seem that the low number of pieces implies that the problem should be easy (at least for most chess players). However, this is one of the top rated Chess Tempo problems, ranked as the 52^{nd} out of 48,710 problems at the time of this writing, with the rating of 2450 rating points (other Chess Tempo statistics of this problem include: 211 users attempted to solve it, spending 602 seconds on average and with success rate of 31.75%).

What makes a particular chess tactical problem difficult? In order to understand it, we must first get acquainted with the solution. The solution of the problem in Fig. 1, shown in standard chess notation, is 1.Rh7-h8+ Ba3-f8 2.Bf6-g7! (2.e6-e7? Ke8-f7!=) Ke8-e7 3. Bg7-h6!! and Black loses in all variations, e.g.: 3... Rc8-a8 4.Rh8-h7+! Ke7-f6 5.Rh7-f7+ and the black bishop is lost. White's 3^{rd} move (3.Bg7-h6!!), virtually giving an extra move to the opponent, is particularly difficult to see in advance. Note that 3.Bg7xf8? Rc8xf8 4.Rh8-h7+ achieves nothing after 4... Ke7-f6!, with a draw. In the present case, it was not only the case that white was required to make the highly unexpected and counterintuitive move 3.Bg7-h6!!, there were also some seemingly promising alternatives that actually fail to win.

B. Hypothesis

Our hypothesis is that one's ability to estimate the difficulty of a problem is positively correlated with his or her expertise and skills in the particular problem domain. In chess, for example, such expertise and skills are usually measured by the World Chess Federation (FIDE) Elo rating. However, we conceive of chess strength as only one among multiple factors influencing the ability to make good predictions. For example, in the case of teaching, one should develop skills related to estimating difficulty in order to select appropriate tasks for one's students. Exhibiting greater expertise in a domain (e.g., being a stronger chess player) should (in principle) increase the chances of making better predictions – due to increased awareness of various possibilities and their potential consequences. However, for a group of people of similar expertise, the problem's difficulty may vary due to their specific knowledge and individual style. Moreover, it is important to note that FIDE Elo rating does not solely reflect chess players' tactical skills, but also their strategic knowledge etc. Hence, we do not necessarily expect a high linear correlation between player's FIDE Elo rating and their success in ranking the positions.

C. Modeling the difficulty of tactical positions

Guid and Bratko [2] proposed an algorithm for estimating the difficulty of chess positions in ordinary chess games. However, we found that this algorithm does not perform well when faced with chess tactical problems. The reason for this is that computer chess programs tend to solve tactical chess problems very quickly, usually already at the shallowest depths of search. The above mentioned algorithm takes into account the differences in computer evaluations when changes in decisions take place with increasing search depth, thus the computer simply recognizes most of the chess tactical problems to be rather easy, and does not distinguish well between positions of different difficulties (as perceived by humans). Estimating difficulty of chess tactical problems therefore requires a different approach, and different algorithms. It is therefore necessary to investigate the way the players of different strength solve tactical problems and estimate their difficulty, and to better understand what may be the properties of such difficulty estimation algorithms. Hence, we have used physiological measures that gauge performance in chess players' ability to assess the difficulty of tactical problems, in

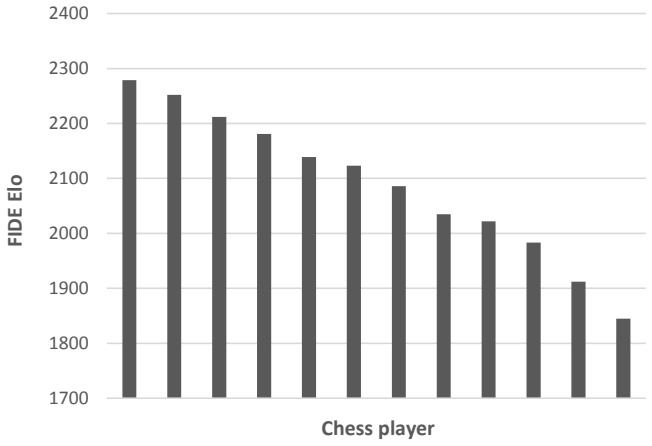


Figure 2. FIDE Elo ratings of the participants.

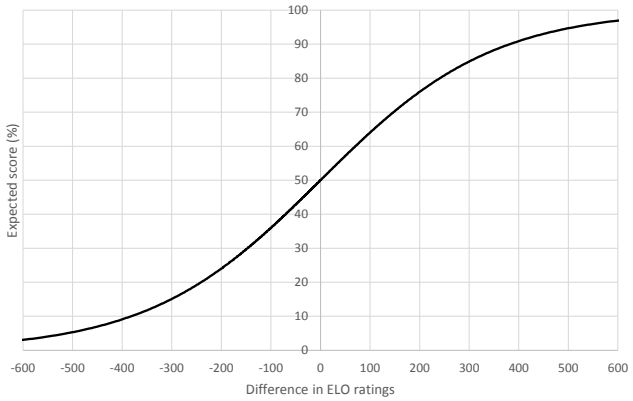


Figure 3. The Elo curve and expected scores.

addition to qualitative reports on perceived difficulty and on variations considered during problem solving.

III. METHODOLOGY

In the experiment, 12 chess experts solved and then ranked a selection of Chess Tempo problems according to their estimated difficulty. Only problems with established difficulty ratings (each attempted by at least 575 Chess Tempo users) were used. The participants consisted of 10 male and 2 female chess players (average age: 48 years). Their FIDE Elo ratings vary between 1845 and 2279 (average: 2089) and are given in Fig. 2. The Elo rating system [23] is adopted by FIDE (World Chess Federation) to estimate the strength of chess players.

Fig. 3 shows the Elo curve, i.e., a plot of the expected score at particular rating differences between two players. It is shown here in order to give the reader an approximate idea about the relative strength of the participants. Assume, for example, that two players are rated r_1 = 2200 and r_2 = 2000. The difference between r_1 and r_2 is 200 rating points in this case. According to the Elo rating system, the expected success rate of the higher rated player playing against the lower rated player is 76% and the expected success rate of the lower rated player is 24%. The expected scores do not depend on the actual ratings r_1 and r_2, but only on their difference. The expected score between two players would also be 76:24 according to the Elo curve if their ratings were, say, r_1 = 2050 and r_2 = 1850, because the rating difference in this case is also 200 points.

Eye tracking was used in order to gather perceptual data about performance and difficulty. One of the main advantages of eye tracking is that there is no appreciable lag between what is fixated and what is processed [24]. The aim was to have a grip on what is happening when the players were solving the problems, in order to understand better why a particular player missed the correct solution, what happened when a particular problem was underestimated, what piece movements did the player focused upon etc. In the experiments, the chess problems were displayed as ChessBase 9.0 generated images, 70 cm from the players' eyes. Participants' head was stabilized by a chin rest. Fig. 4 shows the experimental setting in the eye-tracking room. The players' eye movements were recorded by an *EyeLink 1000* (SR Research) eye tracking device, sampling at 500 Hz. Nine-point calibration was carried out before each part of the experiment session.

Participants were presented with 12 positions – chess tactical problems – randomly selected from Chess Tempo according to their difficulty ratings. Based on their Chess Tempo ratings, the problems can be divided into three classes of difficulty: "easy" (2 problems; their average Chess Tempo rating was 1493.9), "medium" (4; 1878.8), and "hard" (6; 2243.5). While the problems within the same difficulty class have very similar difficulty rating, each of the three classes is separated from the other by at least 350 Chess Tempo rating points. Some problems may have more than one single correct solution. Table I displays the statistics for the 12 tactical chess problems: Chess Tempo rating, success rate and the number of attempts by Chess Tempo users, average problem solving times, the number of correct solutions, and our difficulty class.

The 12 positions were presented in 3 blocks of four positions, randomized within the blocks and between blocks to avoid a sequence effect. There were short breaks to prevent the accumulation of fatigue. The experiment with each player lasted between 20 and 45 minutes. The subjects were instructed to input their solution (their suggested best move) as soon as they have found a winning solution. They were not allowed to exceed the time limit of three minutes for each position.

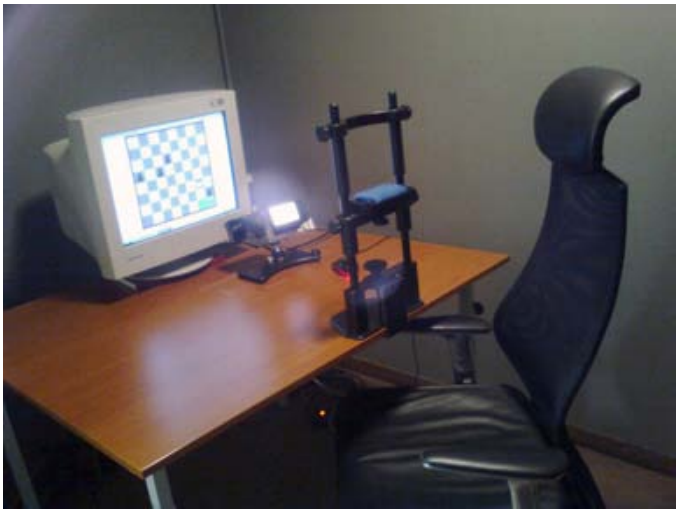


Figure 4. The experimental setting in the eye-tracking room.

TABLE I. CHESS TEMPO STATISTICS OF THE PROBLEM SET.

#	Rating	Success	Attempts	Average time	Solutions	Difficulty
1	1492.5	61%	789	3:50	2	easy
2	1495.3	62%	712	2:12	2	easy
3	1875.2	49%	669	4:08	3	medium
4	1878.1	51%	626	3:31	1	medium
5	1878.6	52%	774	3:16	1	medium
6	1883.3	53%	694	6:39	2	medium
7	2230.9	37%	809	6:53	1	difficult
8	2233.1	36%	815	6:13	1	difficult
9	2237.5	34%	575	7:01	1	difficult
10	2238.5	38%	751	5:20	1	difficult
11	2243.4	40%	572	8:49	1	difficult
12	2274.9	38%	580	9:41	1	difficult

TABLE II. THE PROBLEM-SOLVING STATISTICS.

#	Rating	Success	First moves	Pieces	Avg. time (sec)
1	1492.5	83%	4	3	71.5
2	1495.3	100%	2	2	65.5
3	1875.2	100%	2	2	67.4
4	1878.1	33%	5	3	105.0
5	1878.6	42%	4	3	101.3
6	1883.3	100%	1	1	91.6
7	2230.9	25%	2	2	78.5
8	2233.1	42%	5	3	95.0
9	2237.5	67%	3	2	113.5
10	2238.5	75%	3	2	96.3
11	2243.4	33%	3	1	120.0
12	2274.9	33%	3	1	123.5

Retrospective reports were obtained after the completion of the experiment. These reports serve as a key to understanding the way experts approached the presented position, and to the variations they considered. Chess experts are able to remember variations and are capable of reconstructing even full chess games. Hence, the retrospective reports obtained should have high validity. After the experiment, participants were asked to rate the problems (from 1 to 12) in ascending order of their difficulty. They were *not* told that the problems were divided into three difficulty classes, in order to avoid the bias introduced by this information.

The data types of primary importance to our investigation were: success rate in solving and in ranking the positions, and the type of solutions that players considered (also the incorrect ones). Success rate is an objective parameter, associated with the difficulty of the problem. It shows whether the person was able to solve the problem correctly. In combination with the retrospective reports, it provides an additional framework for understanding participants' estimation of the difficulty of particular problems. On the other hand, the measure of success rate does not account for the way that people went about solving the problem. We analyzed the success rate of the participants in ranking the positions while using Chess Tempo's (well established) difficulty ratings as a frame of reference, in order to observe how good chess players were at estimating the difficulty of problems. We found that in the cases when players did not solve the problem correctly, they tended to make a gross error in their estimate of the difficulty of the position.

The program *DataViewer* was used to generate reports about the participants' eye-gaze activity: saccades, fixations, interest areas, and trial reports. The data analysis will be discussed in the next section.

IV. ANALYSIS OF EXPERIMENTAL RESULTS

A. Statistical Analysis

We computed the correlation between various difficulty rankings for the set of chess positions. The rankings come from individual players that took part in the experiment, and from the Chess Tempo database. The Chess Tempo ranking order was derived from the Chess Tempo difficulty ratings of individual positions (see Table I). The players did not estimate difficulty ratings, but produced their ranking orders directly. That is, they were asked to rank the positions in order: from easiest to most difficult. We used Kendall's tau (τ) rank correlation coefficient which we applied to our data as follows. Given two rankings, Kendall's τ is defined by:

$$\tau = \frac{n_c - n_d}{n * \frac{n-1}{2}} = \frac{n_c - n_d}{n_c + n_d} \quad (1)$$

Here n is the number of all chess positions in the rankings, and n_c and n_d are the numbers of concordant pairs and discordant pairs, respectively. A pair of chess positions is *concordant* if their relative rankings are the same in both ranking orders. That is, if the same position precedes the other one in both rankings. Otherwise the pair is *discordant*. In our data, some of the positions were, according to Chess Tempo, of very similar difficulty. Such positions belong to the same difficulty class. To account for this, the formula above was modified. In the nominator and denominator, we only counted the pairs of positions that belong to different classes.

Table II shows respectively: position numbers and their Chess Tempo ratings (see Table I for more details about the problem positions), the rate of correct solutions by the participants, the number of different first moves tried, the number of different pieces considered for the first move, and the participants' average time spent on the problem.

Fig. 5 shows the relation between Kendall's τ and FIDE Elo ratings for each of the 12 participants. Pearson product-moment correlation coefficient (Pearson's r) was computed in

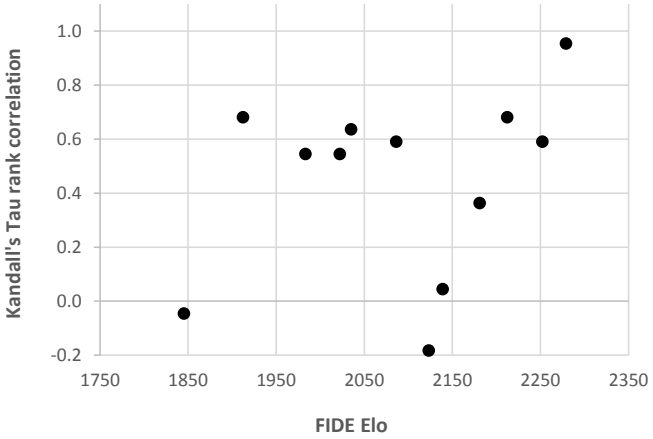


Figure 5. The relation between Kendall's τ and FIDE Elo ratings.

TABLE III. PARTICIPANTS' RESULTS OF PROBLEM SOLVING AND DIFFICULTY ESTIMATION.

Participant #	1	2	3	4	5	6	7	8	9	10	11	12	–
FIDE Elo	*2279*	*2252*	*2212*	*2181*	*2139*	*2123*	*2086*	*2035*	*2022*	*1983*	*1912*	*1845*	*Chess Tempo*
"easiest"	2	3	2	3	3	7	2	2	4	3	8	3	1
	1	2	1	10	2	8	3	3	5	2	2	9	2
	3	6	3	1	12	12	10	4	6	5	11	10	3
	5	1	10	2	5	6	1	7	2	1	7	1	4
	6	10	6	7	6	3	6	1	9	7	1	2	5
	10	4	9	6	4	2	4	10	3	8	6	11	6
	4	7	7	9	1	1	8	5	7	4	9	12	7
	9	12	5	8	10	9	12	9	1	6	5	7	8
	11	9	4	5	7	4	7	6	10	10	12	6	9
	12	8	12	4	9	10	9	12	8	11	10	4	10
	7	5	8	12	8	11	11	8	11	12	3	5	11
"hardest"	8	11	11	11	11	5	5	11	12	9	4	8	12
Discordant pairs	1	9	7	14	21	26	9	8	10	10	7	23	–
Kendall's τ	0.95	0.59	0.68	0.36	0.05	-0.18	0.59	0.64	0.55	0.55	0.68	-0.05	–
Solved correctly	11	8	8	5	7	8	6	8	5	8	9	6	–

order to determine the relationship between Kendall's τ and the chess strength of the participants (reflected by their FIDE Elo rating). There was a moderate positive relationship that is statistically not significant between Kendall's τ and FIDE Elo ratings (r = .30, n = 12, p = 0.34). Clearly, there is no linear correlation between player's Elo rating and their success in ranking the positions.

Table III demonstrates big discrepancies between ChessTempo rating and participants estimation of difficulty. It shows the difficulty rankings each participant gave to the positions they solved. For example, the chess player with FIDE Elo rating of 2279 ranked the positions in the following order: 2 (the easiest one according to the player), 1, 3, 5, 6, 10, 4, 9, 11, 12, 7, 8 (the most difficult one). The "correct" order according to the Chess Tempo ratings is given in the last column of the table. Notice that the numbers of positions refer to the position numbers given in Table I: Positions 1-2 are from the difficulty class *easy*, Positions 3-6 are from the difficulty class *medium*, and Positions 7-12 are from the difficulty class *difficult*.

As it can be seen from the table, on several occasions our participants ranked a position from the class *difficult* to be easier than a position from the class *easy*, and vice versa. Keep in mind that the difficulty classes are clearly separated by more than 350 Chess Tempo rating points. Although Chess Tempo ratings only resemble FIDE ELO ratings (they are not on the same scale), a difference of 350 points – or even 700 points, i.e., the minimal distance between the difficulty classes *easy* and *difficult* – represents a huge difference in difficulty.

We were mainly interested in the number of mistakes made in the comparison of pairs that belong to different difficulty classes, and not the ones within a class. Thus, when computing the value of Kendall's τ, we only counted the pairs of positions that belong to different classes as discordant pairs. The above mentioned player ranked Position no. 2 before Position no. 1, however, this is not a discordant pair, since they both belong to the difficulty class *easy*. The only discordant pair of this player is 10-4, since Position no. 10 is from the difficulty class *difficult* and Position no. 4 is from the difficulty class *medium*. As another example, let us briefly mention discordant pairs by the second-best rated chess player (FIDE Elo 2252): 3-2, 3-1, 6-1, 10-4, 10-5, 7-5, 12-5, 9-5, and 8-5. At the bottom of the table the number of correctly solved problems is displayed for each of the participants.

Chess players obtain their FIDE Elo ratings based on chess tournament games. However, they may not be a reliable predictor of the players' tactical skills. Even the correlation between their FIDE ratings and the performance at solving the experimental problems was surprisingly unclear. In order to verify this, we observed the relation between players' FIDE Elo ratings and the number of correctly solved tactical problems that were the subject of our experiment. The results are demonstrated in Fig. 6. Players' FIDE Elo ratings were rather poor predictors of the players' success in solving the given tactical problems. This is not completely surprising, as chess strength is dependent upon multiple factors in addition to the tactical ability. Nevertheless, this result provides an explanation for why estimating difficulty of chess tactical problems cannot be strongly correlated with players' FIDE Elo ratings. Perhaps Chess Tempo ratings would be a more reliable predictor for this purpose, however, these ratings were unavailable, since several of our participants were not Chess Tempo users.

We then observed the relationship in players' success

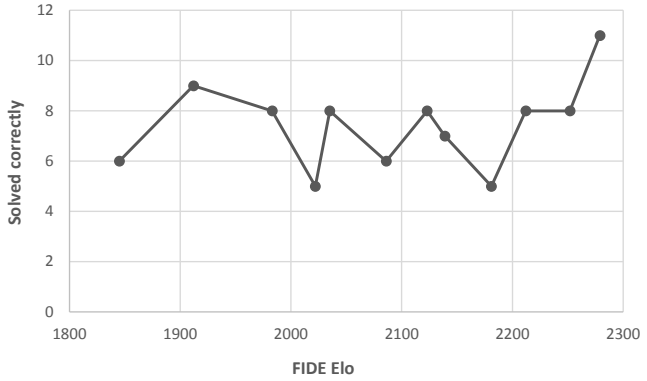


Figure 6. The relation between players' FIDE Elo ratings and their success in solving tactical chess problems.

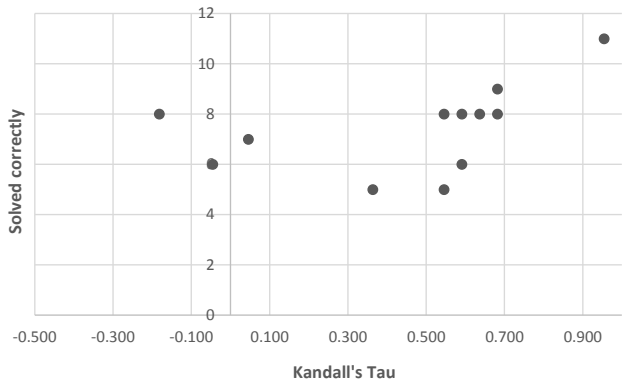


Figure 7. The relation between players' success in estimating difficulty of the problems and their success in solving these problems.

in estimating difficulty of the problems (according to the Kendall's τ rank correlation coefficient) and their success in solving the problems correctly. The results are demonstrated in Fig. 7. There was a moderate positive (statistically not significant) relationship between Kendall's τ and the problem-solving success rate (r = .44, n = 12, p = 0.15). It is interesting to notice that this relationship is slightly stronger than the relationship between Kendall's τ and FIDE Elo ratings (given in Fig. 5), which is in line with the observations stated in the previous paragraph.

Questions remained about the reasons why some rather strong players (according to their FIDE Elo ratings) performed rather poorly at estimating the difficulty of the problems, as well as at solving them correctly (and vice versa). For this purpose, we analyzed the data from the eye-tracking sessions and from the players' retrospective reports. This analysis is the subject of the following section.

B. Eye Tracking

A crucial part of eye tracking data processing is the analysis of fixations and saccades in relation to the squares of the chess-board, defined as interest areas (IAs) [25]. We analyzed what percentage of the fixations fall on a particular interest area: 1) for each individual, 2) for all fixations of all participants. For the purpose of the analysis, the following phases were focused upon: 1) the first 10 seconds after presentation; 2) overall duration of the trial. The first 10 seconds represent the *perceptual phase* according to [26].

De Groot [27] conducted several think-aloud protocols with chess players of different strengths, and discovered that much of what is important to decide on the best move occurs in the player's mind during the first few seconds of exposure to a new position. He noted that *position investigation* always comes before the investigation of possibilities. Furthermore, he divided the initial phase of the thought process into *static*, *dynamic*, and *evaluative investigation*, and found that considering the position from these three points of view typically occurs in this fixed order. Eye movement studies showed that during a few seconds exposure of a chess position, masters and novices differ on several dimensions, such as fixation durations and the number of squares fixated. Retrospective protocols indicated that very little search is conducted during these first few seconds [28].

Fig. 8 demonstrates two *EyeLink* duration-based fixation maps (visualized as "heatmaps") of Position 3. The displayed heatmaps depict the areas upon which two of the participants spent the greatest amount of time looking at. The left-hand diagram depicts the fixations made by Participant 1, and the right-hand diagram the fixations by Participant 4. The FIDE Elo ratings of the two participants are 2279 and 2181, respectively, and the first participant was more successful both in terms of ranking the positions according to their difficulty as well as in solving them correctly (see Table III for details). Position 3 has three possible solutions. The quickest way to win is mate in 4 moves: 1.b3-b4 (avoiding drawing due to stalemate – i.e., when the player to move has no legal move and his king is not in check) a5xb4 2.Kg3-f2 b4-b3 3.Kf2-f1 b3-b2 4.Sh3-f2 checkmate. However, there are two alternative solutions, which begin with the White Knight jumping to squares g5 (1.Nh3-g5) and f2 (1.Nh3-f2+), respectively. In this case, the two motifs (sacrifice a pawn and deliver checkmate vs. merely move the knight to avoid stalemate) are neatly separated on the board so that eye activity can be reliably attributed to each variation.

The heatmaps show that Participant 1 (depicted in the left-side diagram), i.e., the stronger player according to the FIDE Elo ratings, focused upon the quickest path to checkmate, while Participant 2 (see the right-side diagram) looked at the first of the alternative moves. Interestingly, the stronger player correctly assessed this position as the third easiest one, while the other one assessed it as the easiest position of the whole set (see Table III). This may be contributed to a possible message by the two heatmaps: the second player (right-side diagram) most likely did not notice that there exists a quick and effective solution which however demands a sacrifice of a pawn in order to avoid stalemate. It is stalemate in this position that causes some players to go wrong by moving White King to f2 (not noticing that this move results in no legal moves for the opponent), thus contributing to the higher rating of this problem (compared to the lower-rated Positions 1 and Position 2). We briefly note that the stronger player also spent less time on this position (20 seconds vs. 36 seconds).

Fig. 9 shows an alternative type of *EyeLink* fixation map for Position 4 – one of the positions that was regularly estimated by participants to be more difficult than its Chess Tempo rating (1861) indicates. The problem has only one correct solution - attacking Black Queen on b3 with the move 1.Nc2-a1. The retrospective accounts of the variations the players considered indicate the presence of two main motifs that all participants attended to: 1) weakness of Black King on e8; 2) trapping Black Queen on b3. The diagrams from the perceptual phase (see the left-side diagram Fig. 9) and the data from players' retrospective reports confirm that all participants spotted the first motif. The players considered different variations aiming at exploiting this motif (see the solid arrows in the right-side diagram Fig. 9): attacking with Re4xe7 or strengthening their attack through playing Qc1-e3. During the perception phase and for the overall duration of the trial, the e7 square is the most attended IA – accounting for 9.5% of the fixations in perceptual phase, and 9.3% of the fixations in overall duration of the trial, respectively. Another main piece in this motif, Re4, is the third most visited area, accounting for 7.3% of the fixations in the perception phase.

The other salient motif in Position 4 has also been reported

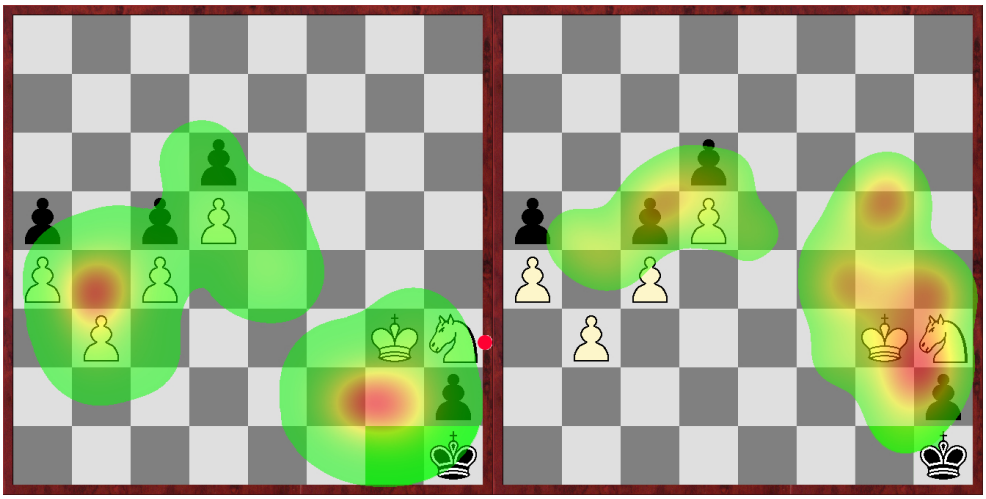


Figure 8. The *EyeLink* fixation maps for Participant 1 (left) and Participant 4 (right), showing the areas that the two players were focused on.

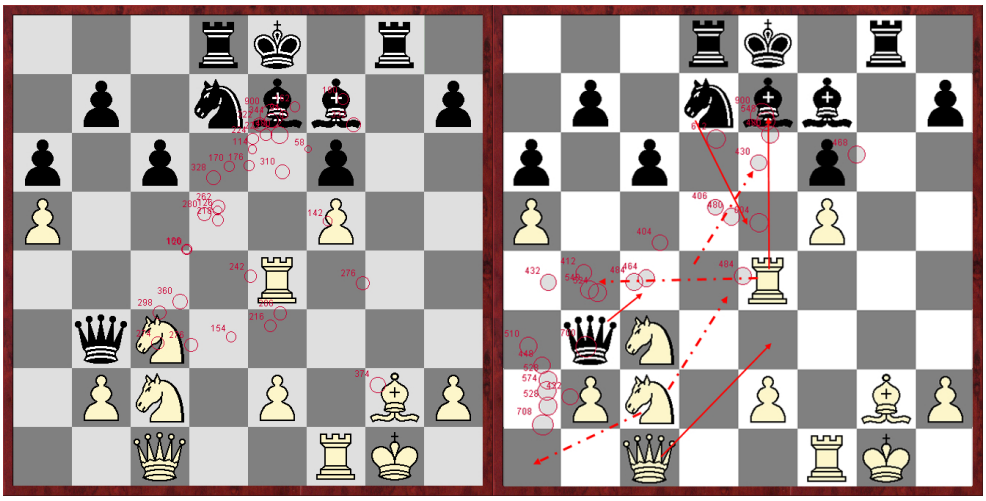


Figure 9. The *EyeLink* fixation maps of a random participant for the first 10 seconds (left) and overall duration of the trial (right), for Position 4.

in the retrospective accounts provided by all participants: trapping Black Queen on b3. As shown on Fig. 9 (right side, see the dashed arrows) three moves were considered by participants: 1.Re4-b4, 1.Nc2-d4 or 1.Nc2-a1. The percentage of fixations recorded on a1 is low – 0.3% of the whole trial. A possible explanation is that once the potentially winning move Nc2-a1 is spotted, the calculations should be focusing on the squares surrounding the Qb3 – to verify whether this move leads to a success in trapping the Queen. Also, the rate of the fixations on a1 may be influenced by the fact that a1 is a corner square. During the perceptual phase the White Knights on c2 (2.9%) and c3 (8.9%) – note that they are both on the squares surrounding the Qb3 – were among the fixations attended to for the longest period of time.

Our data shows that despite their differences in strength, participants' line of thought focused on the above two motifs. This position has only one good solution (1.Nc2-a1), but two salient motifs (two families of branches of the search tree). The first motif triggers variations that do not contain the right solution. It is evident and invites for violent moves in the center of the board and along the e-file. This motif is even more appealing as White has two Knights at her disposal – pieces that are usually strong in the center of the chess board. The candidate moves are: Re4xe7 - direct attack; Qc1-e3 - strengthening White's attack. The second motif's candidate moves appear less intuitive. Choosing to move a Knight to the edge, or even to the corner (a1), is a rather counterintuitive move since Knights are considered to be strongest in the middle of the chessboard. Ultimately, the aforementioned characteristics of the problem create predisposition for increased difficulty even for skilled chess players. Hence, the success rate for this position was 33% only.

The White Knight on c2 was identified as the piece that should be used in the first move of the winning variation in this tactical position by 66% of the participants. However, half of these players were simply unable to see the move 1.Nc2-a1, most likely because all chess players are taught not to move a knight into a corner. Putting the knight on such square reminds chess experts on the well-known expressions like "A knight on the rim is dim" or the French "*Cavalier au bord, cavalier mort*" ("A knight on the edge is dead"). Neiman and Afek [29], who analyzed the reasons why some moves are

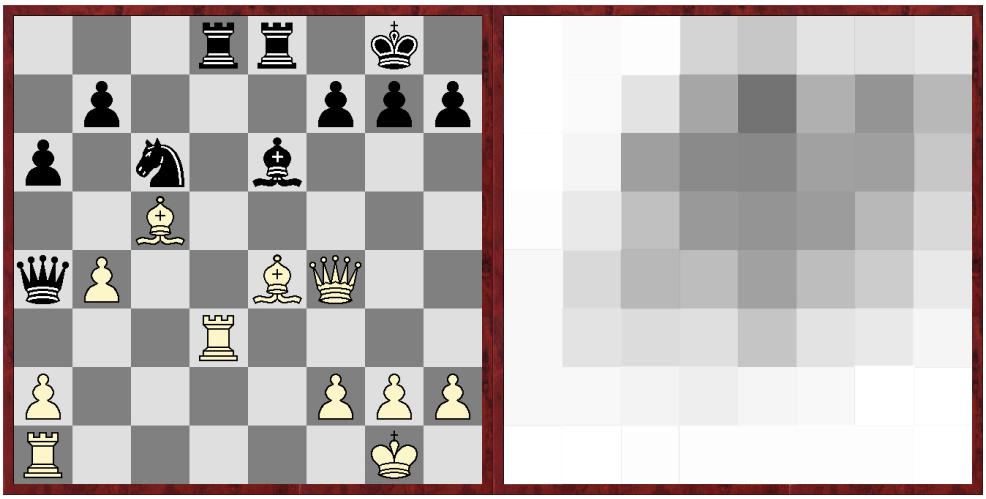


Figure 10. Left: Position 10; right: the *EyeLink* fixation map in this position for overall duration of the trial and averaged across all participants.

often "invisible" to chess players, discovered that amongst all the possible backward moves with the chess pieces, the hardest to spot are those by the knight. Actually, the incorrect alternative 1.Nc2-d4 – putting the night in the center – is so natural that it makes the correct (but backward!) move 1.Nc2-a1 very difficult to find for many players. This is an example of a mistake made due to negative knowledge transfer [30] when the player overlooks the solution of the problem as a result of their training. In other words, seemingly good moves can increase the difficulty of a chess position due to a simple (but misleading) heuristics that people may use in order to solve the problem. A famous example of the negative impact of prior experience is the so-called Einstellung (mental set) effect, which applies to a wide range of problem-solving settings where the presence of a familiar pattern can actually block the discovery of better solutions [31], [32], [33].

Fig. 10 (the left-side diagram) demonstrates Position 10, which was one of the most difficult positions in the experimental set (Table I). However, most of the participants underestimated its difficulty.

The solution is a sequence of moves based on a geometrical motif:

- Step 1: White Queen moves to h4, where it simultaneously attacks both h7 (thus threatening checkmate) and Black Rook on d8.
- Step 2: Black should use the next move to defend against checkmate, thus has no time to protect or move the Rook.
- Step 3: White exchanges the White Bishop for Black Knight (attacking Black Queen at the same time), to remove the crucial defender of Black Rook on d8.
- Step 4: Black should recapture the White Bishop, since Black Queen is under attack.
- Step 5: White wins Black Rook on d8, taking it with White Rook, supported by White Queen.

According to Chess Tempo statistics, about 60% of users failed to solve this problem. In this particular case, good combinatorial vision is required in order to recognize the geometrical pattern. Once the motif is spotted, the solution may seem rather easy. In our experiment 75% of participants solved this problem correctly, which is probably the reason for the underestimation of its difficulty.

On the right side of Fig. 10, the more frequently viewed squares according to the eye tracking data are shaded in darker grey (and vice versa). This information was obtained by averaging the fixation maps of all participants for overall duration over the trial, thus representing the "collective" fixation map for Position 10. It was interesting to observe, also on the basis of individual fixation maps in perceptual phase, that all participants focused on roughly the same part of the board. However, although one would expect that the squares that play the major role in the above presented geometrical motif (such as h4, h7, d8, c6, and e4) would stand out in this diagram, this is not the case. The most viewed square by the participants was e7, which does not play any particular role in the problem solution – except that it is positioned somewhere in the middle of the above mentioned squares. On several occasions – one of them is also the move 1.Nc2-a1! in Position 4, as explained earlier – we spotted that the players found the best move, although they barely looked at the square with the piece that is about to execute it. This reflects some of the limitations of eye tracking research when exploring higher cognitive functions (as in the case of solving chess tactical problems).

One explanation is that eye tracker records the position of the focus of the eye. However, neighboring squares are also visible to the person. In the case of Position 4, the low amount of fixations on a1 may be due to it being a corner square, or just because the player had to calculate the implications of the move Nc2-a1 for the pieces surrounding Black Queen. In both cases, there is no deterministic one-to-one mapping between the physiological data (fixations) and higher cognitive processes. Hence, in our study, the eye-tracking data proved to be most useful when providing physiological evidence of the areas (groups of adjacent squares) on the chess board that people attended to.

Analyzing eye tracking data together with the retrospections provided the basis for the previously described case studies. Eye tracking data enables the verification that a player's retrospection is a genuine account of her thought process when solving the problem, and not a post-hoc justification for her decision. In this way, they can also provide clues about the source of difficulty of a position.

C. Retrospection Reports Analysis

The retrospective reports represent an important source of information for better understanding of how the participants tackled the given problems, and what were the candidate moves and variations they considered. In this section, we briefly analyze what we learned from retrospection analysis of Position 5 (see Fig. 11). This position is an example of a position with many motifs, although they are very unsophisticated. Each motif is actually a direct threat to capture a piece in one move, as shown by the arrows in Fig. 11: both Queens are under attack (Nc6xa5, Rd8xd1, Ne3xd1) and there are many further direct threats to capture pieces (Nc6xd8, Ne3xf1, Ne3xg4, f5xg4). These single-move "motifs" are so straightforward that they hardly deserve to be called motifs due to their conceptual simplicity.

In their retrospections, the players mentioned all or most of the motifs shown in Fig. 11. Even if the motifs themselves are straightforward, the players' typical comment was "a rather complicated position." Only 50% of the players found the only correct solution b7xc6, and the most frequent incorrect solution was Rd8xd1. What makes this position difficult is the large number of simple motifs (threats) which combine in many different ways. This gives rise to relatively complex calculation of possible variations where various subsets of the "motif moves" combine in different orders. In this particular case, this is enough to make a position difficult for a human.

This case very clearly supports the following tentative conclusions indicated by the retrospections concerning other positions as well. First, the retrospections nicely conform to the early model by De Groot [27] of chess players' thinking about best moves in chess. De Groot's model conceptually consists of two stages: (1) positions investigation (in this paper referred to as "identifying motifs"), and (2) investigation of possibilities, or search (here referred to as "calculation of variations"). Strong chess players have to master both of these two tasks. But an interesting question is: which of the two tasks contributes more to the difficulty? The tentative conclusion from our analysis of retrospective reports is that this is task 2, i.e., calculation of variations. At least for players of Elo rating between about 1800 and 2300 (our players' range), the calculation skill seems to be the more important factor. The motifs detected in our positions are almost invariable between the players. The success in solving the positions however varies considerably, which is due to the different strengths at calculation of variations. These differences are not only reflected in the correctness of the solution proposed by the players, but can also be clearly detected in the players' comments that include many mistakes in the calculations.

It was interesting to notice that missing the correct line of reasoning often leads not only to underestimating, but also to overestimating the difficulty of a position. One of the participants, for example, provided the input 1... Bf8-d6?? (incorrect move) as the solution of the tactical problem in Fig. 11. This move not only fails to win, but also loses very quickly to 2.Nc6xa5 Ne3xd1 3.Bg4xf5+ (the move that the player missed, although 3.Bg4xd1 and several other moves also win for White). However, this participant ranked this position as the most difficult of the whole set of 12 positions – although this position is from the difficulty class *medium*, and therefore its Chess Tempo rating is more than 350 points lower than the ratings of 6 positions in the data set. There were actually two

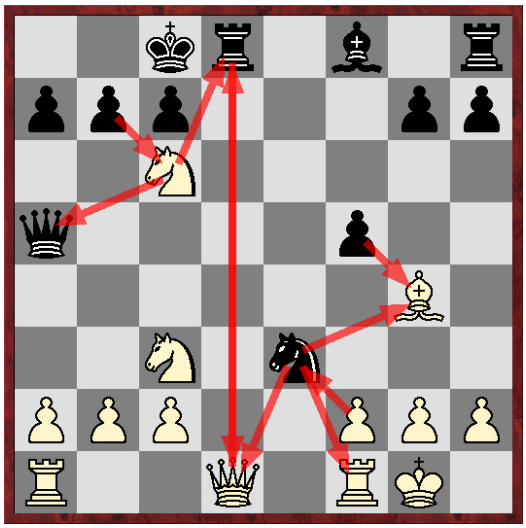


Figure 11. Each arrow indicates a move that corresponds to a separate simple motif, in this case a direct threat to capture an opponent's piece.

participants who labeled this position as the most difficult of all positions in the set (see Table III).

Several participants (5 out of 12) ranked Position 3 (Fig. 8) as the easiest one in the experimental set (refer to Table III). The retrospection analysis revealed that the participants tended to assess this position as a very easy one just because they solved it without much effort after correctly noticing the stalemate motif. However, when assessing the difficulty of such a position, one has to have in mind that not all chess players will notice this motif and that it is likely that many other players may fall into the trap of playing the seemingly logical (but wrong, due to the stalemate) move 1.Kg3-f2, with the idea of putting White King on f1 and then delivering checkmate with White Knight. It is precisely this possibility that caused this problem to score higher, i.e., to obtain a higher Chess Tempo rating. It is interesting to notice that about 50% of Chess Tempo users who attempted to solve this problem failed to solve it correctly.

V. Discussion

As expected, our data indicates that no single measurement directly predicts the difficulty of the task for the players. The best approximation to the difficulty is offered by looking at data such as success rates and solution times.

Difficulty depends on the knowledge of the player and her individual abilities - to spot the most relevant motifs and to calculate concrete variations based on the motifs observed. A tentative conclusion from our retrospection analysis is that the player's strength in calculation of variations is in fact more important than the ability to detect motifs in a position. This seems to be true at least for players in the Elo rating range between 1800 and 2300. This conclusion will be surprising to many since a common view among strong players is that a players' chess strength mainly comes from her deep understanding of chess concepts. The motifs belong to this deep chess knowledge. The calculation of variations is, on the other hand, usually considered as routine activity done without any deep understanding of chess.

Difficulty also depends on the task characteristics, such as the weight of the alternative variations - as this may have an

impact on the degree of uncertainty the player experiences (e.g., the existence of many good or seemingly good solutions may confuse). This is a crucial observation for further attempts to model difficulty.

Regarding the eye tracking data, the analysis of heatmaps and players' retrospections showed that the most attended squares of the heatmap of the player do not necessarily correspond to the squares that the player was thinking about. This is in agreement with general experience in eye tracking research. Instead, a central square of heatmap density should be understood as an indication that the neighboring squares, in addition to the maximal density square, were the specific areas of the players' interest. This is illustrated in Figs. 9 and 10. An interesting future project would be to develop a careful transformation between the heatmaps and the squares on the board that are of genuine interest to the problem solver. Chess knowledge and calculation of variations would certainly be part of such a more subtle algorithm for interpreting eye tracking data.

On the other hand, a potential use of eye tracking data is illustrated by Fig. 8, where the areas on the chess board of the two main motifs were not overlapping. In this and similar cases, the tracking of the player's eye fixations is sufficient to reliably predict what variations are considered.

The players' retrospective reports give important clues on what a mechanized difficulty estimator should look like. It should involve the calculation of chess variations, but not in the way that strong computer chess programs do. The difficulty estimator should carry out a more subtle search guided by the motifs that human players spot in a position. So, only moves relevant to these motifs should be searched, as illustrated in the analysis of the retrospections of Position 4. The complexity of such limited search should eventually produce reliable estimates of difficulty of problems for humans.

VI. Conclusion

The goal of our research is to find a formal measure of difficulty of mental problems for humans. The goal is then to implement such a measure, possibly as an algorithm, which would enable automated difficulty estimates by computers. Obvious applications of this are in intelligent tutoring systems, or in better evaluation of student's exam results, which would take into account the difficulty of exam problems.

In this paper, our study of how to mechanically estimate difficulty was limited to chess problems, more precisely to solving *tactical* chess positions. In solving such problems, humans have to use their knowledge of the domain, including pattern-based perceptual knowledge and the skill of position analysis through calculation of concrete variations of what can happen on the board. Similar kinds of knowledge and skill are required in solving other types of problems, for example in mathematics, everyday planning and decision making, and acting skillfully in unexpected social situations. Therefore, we believe that observations pertaining to difficulty in chess will apply to problem solving in other domains.

Our experiments included observing humans during problem solving (eye tracking, retrospection analysis), and humans themselves estimating the difficulty of problems (ranking of chess positions according to difficulty). One conclusion from this is that estimating difficulty is difficult also for humans, including highly skilled experts. Our experimental results did not confirm statistical significance of the hypothesis that the human's level of expertise correlates strongly with the human's ability to rank problems according to their difficulty. The results in Table III illustrate this point. The players' difficulty rankings of chess problems appear to be almost random!

Also explored was the question of which of the following stages in chess players' thinking about best moves contributes more to the difficulty of chess tactical problem solving: identifying motifs or calculation of variations? The tentative conclusion from our retrospection analysis is that, at least for players of FIDE Elo rating between about 1800 and 2300 (our players' range), the calculation skill seems to be the more important factor in this respect.

In a further analysis of the correlations between the players' rankings and Chess Tempo rankings (considered as the ground truth), and players' Elo chess ratings and the players' success in *solving* the chess problems (not estimating the difficulty), all of these relations turned out not to be statistically significant. The largest correlation coefficient was observed between overall success in difficulty ranking and the overall success in problem solving over all the experimental problems. Although this also turned out not to be statistically significant, it provides an indication that further work in this area may prove to be valuable. Namely, to investigate another hypothesis, i.e., that the success in estimating the difficulty of a particular problem depends on the ability to solve that particular problem.

Acknowledgment

The authors would like to express their gratitude to Kristijan Armeni, Grega Repovš, Anka Slana, and Gregor Geršak for providing support with the preparation of the experiment this study is based on, and to Rob Lee for his comments on an earlier version of the paper.

References

[1] D. Hristova, M. Guid, and I. Bratko, "Toward modeling task difficulty: the case of chess," in COGNITIVE 2014, The Sixth International Conference on Advanced Cognitive Technologies and Applications. IARIA, 2014, pp. 211–214.

[2] M. Guid and I. Bratko, "Search-based estimation of problem difficulty for humans," in Artificial Intelligence in Education, ser. Lecture Notes in Computer Science, H. Lane, K. Yacef, J. Mostow, and P. Pavlik, Eds. Springer, 2013, vol. 7926, pp. 860–863.

[3] B. Woolf, Building Intelligent Interactive Tutors. Morgan Kauffman, New York, 2008.

[4] Y. Wang, Y. Song, and Z. Qu, "Task difficulty modulates electrophysiological correlates of perceptual learning," International Journal of Psychophysiology, vol. 75, 2010, p. 234240.

[5] R. Hunicke, "The case for dynamic difficulty adjustment in games," in Proceedings of the 2005 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology, ser. ACE '05. New York, NY, USA: ACM, 2005, pp. 429–433.

[6] C. Liu, P. Agrawal, N. Sarkar, and S. Chen, "Dynamic difficulty adjustment in computer games through real-time anxiety-based affective feedback," International Journal of Human-Computer Interaction, vol. 25, 2009, pp. 506–529.

[7] W. G. Chase and H. A. Simon, "Perception in chess," Cognitive psychology, vol. 4, no. 1, 1973, pp. 55–81.

[8] P. Chandler and J. Sweller, "Cognitive load theory and the format of instruction," in Cognition and Instruction, L. E. Associates, Ed. Taylor & Francis, 1991, pp. 292–332.

[9] A. D. de Groot, "Perception and memory versus thought: Some old ideas and recent findings," Problem solving, 1966, pp. 19–50.

[10] R.W.Jongman, Het oog van de meester [The eye of the master]. Assen: Van Gorcum, 1968.

[11] H. Simon and W. Chase, "Skill in chess," American Scientist, vol. 61, 1973, pp. 393–403.

[12] E. M. Reingold, N. Charness, M. Pomplun, and D. M. Stampe, "Visual span in expert chess players: Evidence from eye movements," Psychological Science, vol. 12, 2001, pp. 48–55.

[13] F. Gobet, J. Retschitzki, and A. de Voogt, Moves in mind: The psychology of board games. Psychology Press, 2004.

[14] E. Reingold and N. Charness, Perception in chess: Evidence from eye movements, G. Underwood, Ed. Oxford university press, 2005.

[15] F. Gobet and N. Charness, "Expertise in chess," 2006.

[16] K. Kotovsky, J. Hayes, and H. Simon, "Why are some problems hard? Evidence from tower of Hanoi," Cognitive Psychology, vol. 17, no. 2, 1985, pp. 248–294.

[17] K. Kotovsky and H. A. Simon, "What makes some problems really hard: Explorations in the problem space of difficulty," Cognitive Psychology, vol. 22, no. 2, 1990, pp. 143–183.

[18] Z. Pizlo and Z. Li, "Solving combinatorial problems: The 15-puzzle," Memory and Cognition, vol. 33, no. 6, 2005, pp. 1069–1084.

[19] M. Dry, M. Lee, D. Vickers, and P. Hughes, "Human performance on visually presented traveling salesperson problems with varying numbers of nodes," Journal of Problem Solving, vol. 1, no. 1, 2006, pp. 20–32.

[20] P. Jarušek and R. Pelánek, "Difficulty rating of sokoban puzzle," in Proc. of the Fifth Starting AI Researchers' Symposium (STAIRS 2010). IOS Press, 2010, pp. 140–150.

[21] R. Pelánek, "Difficulty rating of sudoku puzzles by a computational model," in Proc. of Florida Artificial Intelligence Research Society Conference (FLAIRS 2011). AAAI Press, 2011, pp. 434–439.

[22] M. E. Glickman, "Parameter estimation in large dynamic paired comparison experiments," Applied Statistics, vol. 48, 1999, pp. 377–394.

[23] A. E. Elo, The rating of chessplayers, past and present. New York: Arco Pub., 1978.

[24] M. A. Just and P. A. Carpenter, "A theory of reading: from eye fixations to comprehension." Psychological review, vol. 87, no. 4, 1980, p. 329.

[25] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer, Eye tracking: A comprehensive guide to methods and measures. Oxford University Press, 2011.

[26] M. Bilalić, P. McLeod, and F. Gobet, "Why good thoughts block better ones: The mechanism of the pernicious einstellung (set) effect," Cognition, vol. 108, no. 3, 2008, pp. 652–661.

[27] A. D. De Groot, Thought and choice in chess. Walter de Gruyter, 1978, vol. 4.

[28] A. D. De Groot, F. Gobet, and R. W. Jongman, Perception and memory in chess: Studies in the heuristics of the professional eye. Van Gorcum & Co, 1996.

[29] E. Neiman and Y. Afek, Invisible Chess Moves: Discover Your Blind Spots and Stop Overlooking Simple Wins. New in Chess, 2011.

[30] R. J. Sternberg, K. Sternberg, and J. S. Mio, Cognitive psychology. Wadsworth/Cengage Learning, 2012.

[31] A. S. Luchins, "Mechanization in problem solving: the effect of Einstellung," Psychological Monographs, vol. 54, 1942.

[32] F. Vallee-Tourangeau, G. Euden, and V. Hearn, "Einstellung defused: interactivity and mental set," Quarterly Journal of Experimental Psychology, vol. 64, no. 10, October 2011, pp. 1889–1895.

[33] H. Sheridan and E. M. Reingold, "The mechanisms and boundary conditions of the einstellung effect in chess: evidence from eye movements," PloS one, vol. 8, no. 10, 2013, p. e75796.

Giving Predictive Abilities to OLAP Systems' Caches

Pedro Marques and Orlando Belo
ALGORITI R&D Centre
Department of Informatics, School of Engineering, University of Minho
PORTUGAL
pcmarkes@gmail.com, obelo@di.uminho.pt

Abstract — **It is not new that on-line analytical processing systems arose to companies to stay. They have the ability to change the most common application scenarios that decision-makers use on their everyday tasks. The large flexibility in data exploration and high performance response levels to queries these systems have make them very useful tools for exploring multidimensional data accordingly to the most diverse analysis perspectives of decision-makers. However, despite all the computational resources and techniques we have today, sometimes, it is very hard to maintain such levels of performance for all application scenarios, analytical systems, or user demands. When context conditions and application requirements change, performance losses may occur. There are a lot of strategies, techniques and mechanism that were designed and developed to avoid (or at least to attenuate) such undesirable low performance situations with the purpose to reduce especially data servers load. On-line analytical processing systems caching is one of them, designed for maintaining previous queries and serving them upon subsequent requests without having to ask the server repeatedly. In this paper, we present an on-line analytical processing systems caching technique with the ability to identify the exploration patterns of its users, i.e., what queries a user will submit during a working session, their frequency and resources involved, and to predict what data they will request in a near future, as well as the sequence of those requests. To do that in an efficient manner, we need to maintain a positive ratio between the time spent to predict and materialize the most relevant views to users, and the time that would be spent if no prediction had been done. Using association rules and Markov chains techniques, we designed a flexible manner to provide an effective caching system for on-line analytical processing systems.**

Keywords – on-line analytical processing; analytical servers; caching; association rules mining; Markov chains; cache content prediction.

I. INTRODUCTION

Due to the amazing increase of companies' data repositories in the last decade, attentions turned to the implementation of more powerful ways of analyzing data. As a consequence, Decision Support Systems, and more specifically, *On-line Analytical Processing* (OLAP) systems [1][2][3][4] are being implemented in a large scale, when compared to what was being done a few years ago. A little everywhere, OLAP and data mining systems have captured the attention of many research teams and creators of large software systems. OLAP systems provide sophisticated mechanisms for the analysis of large volumes of data in a very expeditious way, accordingly to the several exploration perspectives of decision makers. Based on this type of systems, decision-making processes are much more oriented and more effective, being supported by well-structured analytical information and not, as so may times happen, by the simple intuition of a decision-maker supported by a package of statistical data. OLAP mechanisms for data exploration and analysis allow for data to be related in a non-trivial manner, making possible to change the current perspectives of analysis whenever necessary. Thus, they are quite flexible. This is only possible due to the fact that data is stored in very specific structures that were especially designed for this type of analytical processes, faithfully following the multidimensional nature of the data as well as its most regular exploration processes.

The high efficiency of OLAP systems for exploring multidimensional data is based primarily on the pre-materialisation of the data that we believe to be necessary to meet the needs (and sometimes the expectations) of a decision-maker. This pre-materialization process is done recurring to the use of materialized views that potentially allow for satisfying "immediately" any question (query) that is launched to the system by its users. As this ideal case of querying satisfaction is practically impossible to achieve, due in part to the limitations of memory and processing capacity of the systems, several techniques have been developed to improve how these views can be materialized *a priori*. Caching techniques are one of them. For a long time, they were applied in Web systems with very positive results. Since the beginning of its implementation, caching techniques were seen as a way to accelerate the process for responding requests (queries) posted by users. The implementation of a caching system in a Web platform aims mainly to maintain the information that has static properties,

in order to provide it through a cache (and not through the primary data source) to users that request it repeatedly. This gives us two important advantages. First of all, it allows for a great reduction in responses *time-to-user* – a caching server usually is "closer" to the user than the main information source. On the other hand, this avoids repetitive accesses to the main server, freeing it to answer requests that only it can compute.

In OLAP systems, the information to keep data in cache is very dynamic, but it is not updated very often (perhaps only once in each refreshing cycle of their data structures) and when it happens that is usually done in an incremental manner. The challenges posed by all these constraints are not easy to solve, and this is where caching can help. It is already considered as one of the key factors that contributes to a significant part of the improvement in performance of any OLAP system. As an OLAP system evolves, so must the caching system associated to it – Figure 1 illustrates a simple view of a basic caching system for a multidimensional database.

As in Web systems, caching systems were firstly centred on the individual perspective of the user that was the only one to benefit from its own caching architecture (client-side caching). Businesses were quick to realize that this somewhat less-sharing way of caching was not the best approach. Later, the development of caching systems followed the strand of sharing caches between users, whether they were available on the server side or on the customer side.

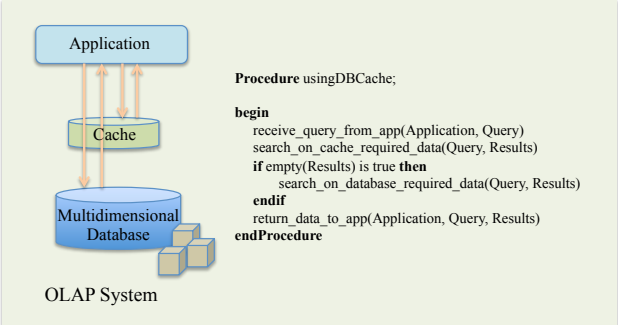


Figure 1. Illustration of a basic caching system for an OLAP system.

As we know, one of the great advantages of OLAP systems is the fact that they can cope with large volumes of data, and execute *ad-hoc* queries within various analysis perspectives giving to decision makers an exceptional way to get more structured insights about company's data. OLAP systems were so well accepted by decision makers that they soon started loading more and more data into them and issuing more complex queries, which quickly surfaced some critical performance issues. As fast as an OLAP Server could be, there is always some space to apply new optimization strategies, trying to improve OLAP servers' performance and OLAP users' satisfaction. Thus, the usage of caching mechanisms in OLAP platforms is a natural (and viable) technological choice when one is concerned to improve the quality of service of an OLAP platform.

Despite being widely implemented and tested, conventional caching mechanisms were not prepared to handle OLAP data. One of the reasons why this type of information was not ideal for caching was due to its dynamic nature (i.e., versus the static nature of HTML information where caching techniques have a particularly good fit). Other aspect to be considered when dealing with OLAP data is the dimension of the data to be kept in cache, both in terms of volume of data as well as in terms of data structure complexity. Comparing again with HTML data, which represents a little effort in terms of space needed to keep it in cache, OLAP data requires a great amount of space, simply due to the fact that any response to a typical MDX (*Multidimensional Expression*) query involves a lot of data, usually materialized in a multidimensional data view (a data cube). Even with the diversity of the data to be maintained, several techniques were developed to apply caching mechanisms to OLAP data [5][6][7], revealing benefits good enough to keep the focus on improving caching techniques in order to integrate them effectively on OLAP server systems.

The work we developed was based on an analysis of today's caching mechanisms and their application in the OLAP field, and based on selected information about user's querying patterns [1]. In order to obtain these patterns, OLAP server logs were fetched, analysed and mined in order to obtain a set of association rules that represent the actions (and consequences) of user's queries (usage profiles), providing us the means to predict future user's querying tendencies. Such predictions unlock the possibility of issuing a query even before a user post it, putting it in cache and finally providing it faster than if no cache was available in the OLAP platform.

Basically, having the ability to establish the exploration patterns of a community of OLAP users give us the possibility to define *a priori* the contents of a cache with a satisfactory confidence label. With this, it will be possible to have, in advance, a predefined set of materialize views that correspond to the most frequent multidimensional queries that were done previously during a certain period. Of course that the success of this strategy depends a lot from the tendencies (and routines) of data exploration that users may have. However, in general terms, an OLAP user uses to apply systematically for a particular set of queries as the starting point of his OLAP session. Knowing that, and having also an exploration profile, we can establish the initial state of a cache, adjusting it dynamically during the execution of an OLAP session accordingly the prediction we have made based on past OLAP sessions - more adequate caches' contents in a shorter time. In short, this was the goal of our work.

In the next sections, a more in-depth analysis to this process will be conducted, explaining the various stages reached along the evolution of the work, as well as

discussing some of most relevant considerations needed to understand the complexity of predicting the multidimensional content of a cache for a specific OLAP platform. This paper is organized as follows: Section II presents a brief relation about some caching techniques and caching maintenance algorithms; Section III shows a detailed overview about OLAP caching, its advantages and disadvantages; Section IV presents and discusses the major characteristics of problems we could face when using high specialized caches as OLAP caches; Section V reveals our technique for a new model of OLAP caching; Section VI reveals and discusses the results of the tests we have done for validating the proposed OLAP caching model; and finally, Section VII presents final remarks and conclusions, as well as point out some lines for future research.

II. Related Work

The introduction of caching mechanisms in OLAP systems brings some great advantages. Firstly, because queries are answered directly from the cache system, decreasing accesses to the OLAP server. Secondly, since caches are usually closer to the users, the network traffic between the nodes closer to the OLAP server decreases - providing better network latency and quality of service. As a result, OLAP server availability and performance increase.

As we know, the concern of improving services of a server, and in particular of an OLAP server, is not new. Many researchers have developed effective work in this area, with particular emphasis in areas such as cache management algorithms and caching systems architectures. Usually, algorithms for managing caches are used to decide whether a given piece of data should be placed (or not) in a cache. This allocation decision affects the contents of the cache and consequently its own performance. Thus, by analyzing the consequences of these decisions positively or negatively it is possible to make an evaluation about the quality of the cache management algorithm that was used, applying a set of metrics especially defined for this purpose, namely the *Hit Ratio* or the *Byte Hit Ratio* metrics.

Over the years many proposals presented a large diversity of algorithms for managing caches. On this domain we should reveal the work that led to the emergence of algorithms such as the *First In First Out* (FIFO), the *Least Frequently Used* (LFU) [8], or the *Least Recently Use* (LRU) [9] algorithms. The operating way of these algorithms is very similar. They are regulated through the definition of static criteria that defines the way as data is removed from a cache when the cache gets full. Independently from the different implementations of data structures that exist to support queues (FIFO), determining the frequency of access of each of the elements maintaining in a cache (LFU) or to sustain a specific time label relative to the time at which data was accessed (LRU), data elements are removed from a cache without concerning their utility for the users of the system.

FIFO, LRU and LFU still are today some of the most popular and theoretically important algorithms for caching management as well as the algorithms *Least Recently Used Second-to-Last Request* (LRU-2) [10], an evolution of the LRU algorithm that was developed to be used in database disk buffering. Finally, a brief reference to a last caching management algorithm: ARC (Adaptive Replacement Cache) [11], which has the ability to balance in adaptive manner the workload of a cache in a self-tuning fashion.

All algorithms for managing caches mentioned earlier aim to manage the information that should be added or removed from a cache system. Similarly, with another level of abstraction, we can refer other works in this domain that address many relevant aspects in the implementation of a caching system, namely the problem of the location of the cache. Alternatively to the implementation of a caching system on the client side, we can do it on the server side, as already referred, creating mechanisms that benefit all users of a given community. From the simple to more complex caching systems, this last approach involves, among others, peer-to-peer, active caching, or chunks based caching architectures. Chunks were defined in Deshpande et al. [12] and are a new indivisible unit. This data unit, with a low granularity level, is mapped in the cache in order to be aggregated to satisfy user requests. The mapping occurs in the server and denotes the relationship between a chunk and the basic units stored in the OLAP Server, allowing for the complementary fetching of data from a main data source.

Let us now look at other approaches, starting by the peer-to-peer architecture. In [5] it was said that, a little bit like as all other proposals in the area of distributed caching, if all participants in a network share their personal caches, everyone would benefit, and proposed a network architecture that enables such features – the PeerOLAP. Additionally, they also presented some other policies for renovating and maintaining data in distributed caching system. In [13] and [14] it was approached an active caching system technique. This technique was presented as an effective solution to the problem of creating a caching system for dynamic information manipulated by Web proxies. As we know, performing caching of static HTML pages is a common practice. But the implementation of a caching system for OLAP queries (and correspondent results) in Web proxies is not very usual. This is due to the fact that proxies usually are not prepared to maintain dynamic information, and especially because they do not have the necessary mechanisms to deal with the necessary post-calculations. On the other hand, in [15] there is a proposal to by-pass problems at the level of caching queries, allowing for the use of data in cache to respond partially to a query. The rest of the answer will be obtained directly from the server. This was accomplished by dividing data into chunks stored directly in the cache. When the level of aggregation of a chunk is lower than of the query, chunks could be used to partially calculate the results of the queries presented. Through the use of mapping mechanisms,

between the data of the query and corresponding chunk number, it was possible to determine all the chunks needed for the calculation of the solution of a given query.

Later, in [16] were proposed other algorithms to group proxies dynamically in "neighbourhoods", regrouping them whenever necessary according some predefined requisites. This approach can be regarded as second level caching, at which information refers to maintain the best neighbours of a certain proxy. The process is relatively simple. When a proxy discovers that there is another one that is not his neighbour but it can bring greater benefits, it adds it to its list of neighbours, eliminating any other proxy in that list that is less beneficial, if necessary – the complexity associated to this process relies mostly in maintaining accurate statistics about the other proxies behaviour and performance.

In OLAP, selecting views to materialize is an NP-complete problem [17]. Some of well known approaches - e.g., [18], [19] or [20] - propose this selection to be performed statically before each set of queries, being results used to respond to subsequent queries. To avoid the problem that arises with the fact that the usage patterns are dynamic, in [21], [22] and [7] it was developed some other techniques to exploit this kind of situations. An evolution of the proposal of the system Dynamat [21] was the implementation of a mechanism regarding the usage patterns of users as well as its dynamic structure [23]. In the same line of research we find the system PROMISE [22], which has the ability to predict in a more accurate manner what was the structure of a query based on some previous usage patterns, as well as the current query issued.

Any proposal that intends to respond to the fundamental problems of a dynamic selection of views – e.g., what is the amount of information that is required to draw a good user profile, and what is the right time to bring such views to memory (only when requested or trying to predict what is the next view to be required), can be found in [23]. As we can see a lot of proposals to design and implement caching systems were done during the last decade. Here, we just enumerated some of them, trying to enhance some important issues about some techniques to put and manage data in a cache. All these issues can be exploited and adjusted for the implementation of specific caches for OLAP systems.

III. To Cache or not to Cache

To cache or not to cache is not a simple decision. The implementation of a caching system in an inappropriate time entails additional costs and does not bring any benefits to the entire system. There are many aspects that must be considered before deciding on the implementation of a caching system. Many of these aspects are related, directly or not, with the existence of a performance problem. To detect or prove it, we can use, for example, profiling or logging techniques that reveal us how the system is being exploited and respond to the information requests. With this, we can find what the information that is most often used is and make sure the system presents it expeditiously. If the system is unable, for performance reasons, to quickly deliver this information, we can improve the system's performance by placing this information in a caching system, which will reduce the number of disk accesses, decrease querying processing time and, consequently, decrease the overall time to get querying results. Additionally, through caching, one achieves the basis to have a more scalable and flexible system, with high service availability and better performance.

Usually, a database system can make three types of caching, namely results, execution plans and data objects. Although they are all important, in our case, we only approached the caching of results, by studying the application of some profiling techniques to querying processing of a given OLAP system. However, whatever the specific area of implementation could be, when implementing caching mechanisms one has to remember that the space available for storing the cache is not unlimited. As a direct consequence we need to choose (and evaluate) what data should be kept (or not) in a cache and what data should be removed giving space for new (and hopefully more relevant) data to the users' needs. Keeping this in mind, researchers started to test quite well known algorithms – frequently referred as cache management algorithms – that up to that time had only been used in other types of environments such as for caching HTML pages with great success. As results became known, there was a clear notion that there should be promoted some additional efforts to develop new breads of algorithms that focused OLAP scenarios in particular.

A caching management system is a crucial element in the overall performance of a caching system. Basically, its main function is to decide on which information must be maintained (or removed) from a cache in order to allow the addition of new data when necessary. With the aim of measuring the performance of such a system, there are two basic metrics, the *Hit Ratio* and the *Byte Hit Ratio*. The *Hit Ratio* is one of the most common ways of evaluating the value of any caching algorithm. This metric is the ratio between the number of requests that were in cache and the total number of requests that were made, and can be calculated using the following expression:

$$Hit\ Ratio \leftarrow RequestsSatisfiedByCache\ /\ TotalRequests$$

However, *Hit Ratio* is not a perfect metric. For instance, even with a higher *Hit Ratio*, the number of bytes served directly by the cache could be smaller than a cache with a lower *Hit Ratio*, which led to the creation of another metric: the *Byte Hit Ratio*. This last metric has been vastly used to evaluate how a cache can satisfy its clients' requests. Contrary to the previous metric, this one is intended to take account not only how many requests were satisfied from the cache, but also how much information was served this way. If the scenario is caching small pieces of data, it is natural

that the percentage of requests answered directly from it is high, but it is also possible that the number of bytes of information satisfied in this way can be low. Conversely, it is possible that a small portion of large pieces of data results in a reverse scenario. The *Byte Hit Ratio* metric is defined according to the following expression:

Byte Hit Ratio ← *BytesSatisfiedByCache / TotalBytes*

As a user of any OLAP (or other) system launches his queries, the cache management algorithm has to check if the necessary information is stored in the cache or. If not, it needs to decide whether it should or should not be added to the cache. If the request cannot be satisfied directly from the cache, there are two possible outcomes:

1) the cache still has space to accommodate the new data, and so it is added without further due, or
2) the cache does not have enough space to store the new data.

In the former case, the content is added, and after that time, when it is requested, it will be served from cache instead of being satisfied directly by the OLAP Server. If there is no space available in the cache management system, the algorithm can either discard this information or free some space in cache in order to add this new data. This is the main decision that cache algorithms have to make. As we know, this decision will affect the way a cache behaves in the presence of new information to be added. One of the most basic ways to do this selection is to use a FIFO approach, which means that the oldest record to have been added to cache will be removed in order to create space for a new entry. If this is not enough, the second (the third, and so on) oldest records will be removed as necessary, record by record. The main problem with this technique is the fact that it does not consider the nature of the data. Despite of its size or actuality, data has an intrinsic value that cannot be measured as simplistically as these approaches propose. Other (more sophisticated) decision metrics were developed using a timestamp of the last access to a specific piece of data [9], the frequency of access to the data [8], or other more complex metadata such as the ones used by the Greedy Dual algorithm [24], for instance. All these metrics, in one way or another, take into account the intrinsic value of data and the relevance each piece of data has to the users and, therefore, they are much more suited to do the (caching) job correctly than others that simply look at the characteristics of the data neglecting its nature and its relevance to users.

IV. OLAP CACHING

Some of the most common operations performed when querying an OLAP server are the well-known drill-down and roll-up operations. The first of these two operations consists of lowering the grain at which the data is being analysed. For instance, we can go down in a hierarchy, detailing systematically, level by level, the grain of the data, from a country-level view to a district-level one, for instance. The roll-up operation is its direct counterpart, allowing viewing data at a higher level following as well a determined hierarchy.

In an OLAP server, the data is stored at the lowest level of granularity and then aggregated to a level required by a specific multidimensional request. In [5] a solution was proposed where this characteristic is explored, mainly by sharing the cache over several cache servers, specifically *OLAP Cache Servers* (OCS). In this approach, each OCS has the capability to apply transformations (aggregations and other operations) to multidimensional structures, and thus combine them to satisfy at least part of a request that has been launched by a user. This way, whenever a user issues a query, the various OCS are asked if they have the needed information and, even if they do not, they are asked again if they can compute it from the data they have at a lower grain than the user requested. This means that an OCS can satisfy not only requests that have been issued before (and cached) but also other issues that involve computations over the data that exists in the OCS.

When configuring an OCS is important to indicate what is the granularity of the data that you want to maintain in a cache, as this will define the type of applications that can satisfy a specific peer. Physically, the data is stored in secondary memory and only brought into primary memory when required or, more accurately, when the fetching algorithms decide the most appropriated time to do that. This operation does not have to necessarily be on-demand and can follow, for instance, some kind of predictive approach [22] or any another technique for fetching data. Another alternative was proposed in [6], where individual caches of users are shared through a peer-to-peer network created between users of a same OLAP System – PeerOLAP. Essentially, this approach was based on the Piazza System [25], and intended to allow a very high level of autonomy in the cache network due to the dynamic nature of Peer-to-Peer networks, where users can connect and disconnect without significantly affecting the overall usability and performance of the system.

As in other proposals following a decentralized architecture, a challenge that this system often faces is the need to establish a mechanism to avoid the uncontrolled spread of messages, which can, as we know, create congestion in the network, deteriorating the overall system performance. A possible solution to overcome this is the definition of a maximum number of "jumps" that a message could give before lose their validity. This is quite intuitive. A message after a given number of “jumps”, even if it can find a peer that has an adequate response, will be hardly provided in the best possible time. Another problem that arises here is when a message is being resent to other users, even before reaching the maximum number of jumps, and the only place where this can be resubmitted it is the data warehouse itself. In this case, the message is not relayed to the central peer such as this could lead to the repetition of

messages sent to the data warehouse, which breaks any kind of goal of a caching system.

As mentioned before, OLAP data is quite dynamic by nature, which means that it is very difficult to predict when the cached data will become out-dated. To deal with this problem, an active caching technique was created [25]. It consists of keeping in the cache server a Java applet that is invoked every time a cache hit occurs. This applet has the role to check with the OLAP Server if the cache information stills valid or if it has changed since the last time it was requested by one or more users. If data stills valid, it will be returned to the user who requested it. If not, the full request will be redirected to the OLAP Server.

One other question that was frequently placed by researchers, was focused on what would be the optimal level of granularity to store data in a cache, in order to not only be able to aggregate it as needed, but also to be able to do that in a timely fashion manner. On such units is the previously presented chunk, defined by Deshpande et al. [12]. When a cache server receives a request from a user, it calculates the parts of that request that it can be satisfied accessing directly the cache, and the information that it need to be requested to the OLAP Server (at a low level of granularity). When all the required data is located in the cache server, it combines it and sends the results to the user, without him ever knowing if the information came from the central server or the cache server. As a last reference we selected the work presented by Sapia [22], which is an approach particularly interesting to us. In that work, the author proposed a predictive system for user behaviour in multidimensional information system environments that explore characteristic patterns users use to show when explore multidimensional data structures. It is an OLAP caching approach that complements other techniques, such as the ones presented in [26] or [15].

Finally, we should say that the maintenance of caches is something that must be included in the routine of any OLAP system administrator. One cannot optimize performance of such a system simply deciding, from one day to another, the implementation of some kind of caching mechanisms as a solution for a current optimization problem. Generally speaking, implementing a caching system by itself cannot solve any optimization problem. Frequently such problems are treated as early as possible, just starting in the querying design phase and evolving their treatment throughout the design chain of a query. In some sense, caches help to solve (or mitigate) such situations. In our case, we were concerned researching some of the most relevant aspects in the maintenance of a caching system for analytical servers, seeing how we could establish a way to "guess" the various forms of data querying that a specific user community practiced. Basically, the idea was only to find a way to characterize their querying patterns, establishing the most used sequences of queries used, and based on that knowledge materialize their results (when possible) in a caching system. And that is what we will explore in the next section.

V. A New OLAP Caching Approach

By their nature, OLAP systems allow for identifying, for each user or group of users, how they access and explore data cubes. If we take the example of a high-level decision-maker, most likely he will access only information regarding to the sales of a specific store or a particular region, avoiding specific and detailed information relating to the sales of all company's products, for example. Exploring features like this, we can define not only which areas of impact in terms of data analysis a user usually does, but also which specific sequence of searches usually he uses to follow. Thus, it is possible to make predictions about what will be the next query sent to the OLAP server by a given user, simply knowing which of the queries he released in the past when doing some data exploration over a set of data cubes. One way to acquire knowledge about the behaviour of a user passes is, for instance, analyzing the log files of previous sessions of data querying. Through the data stored in these files and with the application of specific domain-oriented data mining techniques, it is possible to extract some useful and accurate knowledge about the user usage profile.

One of the issues related to the use of this type of knowledge is his assertiveness and the advantage that comes from its use. Unlike other caching techniques, this approach does not aim to reduce the workload of a data server but to improve the response time satisfying user requests. This type of technique uses the last query launched by a user, and based on it (and other historical data) tries to infer which will be requested next by the user. If it is possible to carry out this prediction, with a sufficiently high degree of certainty, the query and the answer will be immediately placed in the cache so that when the user presents the query its answer will be already stored in memory and the system only needs to provide the results to the user, almost immediately. Another approach involves not only the last query performed, but also a certain number of queries before that. Thus, keeping information about the sequence of requests (queries) made by users it is possible to make predictions with greater certainty. However, you must also have a larger amount of information related to the past behaviour of a user in order to enable the accomplishment of such predictions.

This work was based on the assumption that OLAP system's users have predictable patterns of data that they use to consult on their regular OLAP sessions. The nature of most OLAP users in a company – decision makers – usually means they are focused in a relatively small subset of the data stored in a data warehouse. The day-to-day activity of a decision maker may begin with an analysis of a pre-defined dashboard or an interactive report, and based on the information gathered from the analysis of the data, he will continue his exploration in a lower level view of the same data – probably appealing to a typical drill-down operation. This shows us that for any given user his behaviour will be

repeated during a certain period of time, revealing then a regular usage pattern.

One possible way of extracting these patterns is by analyzing OLAP Server's logs that contain information about what multidimensional queries users had submitted and when they happened. It is also possible to know, for a given user, the sequence of queries he launched between his login and his logout in a specific OLAP session. From the analysis of this kind of information, given a certain period of an OLAP system exploration, another problem arose: how far back in the logs should we go to make sure that the retrieved rules are truly representative of the user's exploration patterns? On one hand, if we analyse the OLAP exploration habits (and tendencies) for a short period of time, we may get rules that represent the most recent patterns and not what the user usually does in the "long run". However, on the other hand, if we analyse a larger period, we may extract rules that represent older OLAP exploration patterns that do not represent what users are doing currently (users may change their exploration habits due to a large variety of reasons, demanding that the algorithm should be able to adapt to such changes).

Taking these constraints into consideration, we began our approach by retrieving the OLAP server's log files, preparing them to be analysed latter by a specific data mining algorithm with the ability to generate a set of association rules that represent the most relevant exploration user patterns – we designate a set of usage patterns by an OLAP profile. From the OLAP server's log files we extract all the MDX queries that were launched during a certain period by a community of users that we want to establish the correspondent data exploration profiles. Each MDX query is fragmented accordingly several dimensions of analysis, such as OLAP session, cube, query, data and time, dimensions, measures, and users. Then, this information is stored in a specific relational data mart that will provide on the next phase the data to data mining association algorithms.

To establish the association rules we used the well-known *Apriori* algorithm [27]. This is one of the most used algorithm for mining frequent item sets, having prove its effectiveness so many times analysing a set of transactions and surfaces the relationships between them, given a minimum value for support and confidence. As it is well known, association rules are usually represented in the format: $A \rightarrow B$ $(sup=\alpha; conf=\beta)$, where *sup* and *conf* represent, respectively, the support and the confidence values of a rule. From an association rule (and from its support and confidence values) we can retrieve two important things, namely the:

- *support* (*sup*), that represents the ratio between the number of times that a sequence of queries A followed by a sequence of queries B was found in the dataset and the total number of queries in that dataset:

$$sup(A \rightarrow B) = \frac{\#(A\ followed\ by\ B\ in\ the\ dataset)}{\#(queries \text{ in the dataset})}$$

- *confidence* (*conf*), that represents the number of times a sequence of queries A is followed by a sequence of queries B in the dataset, divided by the number of times a query A (independently of what query followed it) was found in the same dataset:

$$\text{conf}(A \rightarrow B) = \frac{\#(A\ followed\ by\ B\ in\ the\ dataset)}{\#(A\ in\ the\ dataset)}$$

If we take the association rule $A \rightarrow B$ $(sup=0.3; conf=0.8)$, as a working example, we can say that for every time a user issues the query A he will, in 80% of the cases, issue the query B right after that. On the other hand, we can say that for the analysed dataset, a sequence of queries A followed by a sequence of queries B occurred in 30% of all cases. However, the antecedent (A) of such rules does not correspond necessarily to a single event, which means that A can also represent a set of queries. In our scenario, the prediction process will be supported not only by a single query, but also by a sequence of queries that a user triggered from the beginning of its working session. If the association rule is something like $A1, A2, A3 \rightarrow B$, it means that the consequent of the rule (B) can be predicted as a consequence of the occurrence of the events A1, A2 and A3, with a given confidence and support. These rules allow for more than a simple prediction, "step by step", which may happen when we want to predict what the next query to be executed is. Thus, it is possible to predict *a priori*, with greater anticipation, which queries will be launched by users until the end of their working sessions, as well as to know the sequence those queries.

It can also happen that the antecedents of the rules are not necessarily sets of queries that were issued, but other types of querying conditions environments, like periods of a day, periods of a week, or even business data like sales or stock information. If we explore all these possibilities, associating them to a specific OLAP environment, it is possible to define some specific rules that indicate the frequency of a particular user performing a query, which predictably will be executed at the same time on a given day or week. For example, we can think in an application scenario where a decision agent every Friday afternoon, before leaving its workplace, check systematically the most relevant management indicators in order to get a last view of the business status of the company. On such scenario and time frame, the indicators he analysed were always the same. Nevertheless, the consequences of the analysis will be clarified by a particular set of query that he will launch to verify a particular business case brought to his attention.

Using this type of prediction system reveals some interesting capabilities that can be very useful improving the performance of an OLAP system. However, it also raises

some pertinent questions that should be answered according to the application context of each specific case, namely:

- What is the number of searches of a user that should be taken into account during a prediction process within the same session?
- All rules should be accepted as valid or we need to define some minimum values for support and confidence, from which the materialization of the corresponding views will be rewarded?
- Shall we materialize immediately all queries that we predict will become necessary or only a part of them?

This technique allows us to establish probabilities for the sequence of queries that a user will issue between the beginning and the end of an OLAP session. With this information some actions may be taken to improve the OLAP server's response time to queries. Our approach was to simulate the user's interaction and place in cache the views our algorithm predicted would be used. The main problem with this is the high value of rules that are going to be generated. This could easily produce untreatable results.

Keeping this problem in mind, our work focused on reducing the number of queries that should be included in the prediction phase, without affecting results significantly. To do this, we chose to map all the sequences of queries predicted by the mining algorithm, representing them in a Markov chain [28] as a way to provide a better visual insight of the entire set of generated rules. Next, we defined the minimum value for the confidence associated with the rules that should be used in the prediction phase (*minconf*). Shortly, we discovered that this action would not be enough if we wanted to effectively reduce the number of predicted queries. We needed to optimize the process.

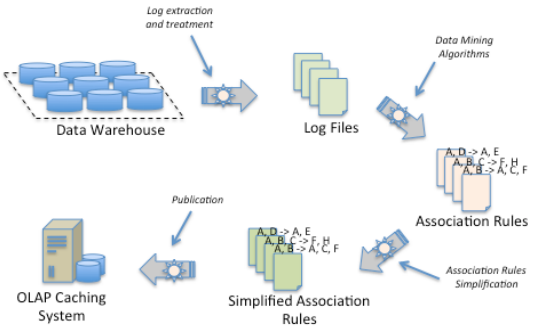


Figure 2. A query sequence prediction for the first dataset.

When removing the rules with a confidence value smaller than *minconf*, we realized that some rules remained without the possibility to be predicted as a sequence of any other query. If we think of the sequence of queries as a graph, and we start removing some of the nodes, there are some of them that lose their entrance arches. Those "nodes" represent the queries that were removed in this second optimization step. This way we also risk an increased number of cache misses, but provide us an alternative way of reducing the number of views to be pre-materialized in the cache. The process followed to establish the set of association rules for a particular *minconf* is depicted in Figure 2.

VI. VALIDATING THE PROPOSED TECHNIQUE

In order to test the technique proposed here, we decided to promote two different test cases, considering the number of query hits achieved before and after the proposed optimization scenarios, for a given set of artificial queries (generated by artificial processing algorithms, not representing the actual usage of an OLAP Server). In Figure 3, we can see the sequence of queries in a Markov chain, which were predicted by the mining algorithm that was used – S_0 and S_8 represent, respectively, the beginning of the session provoked by the user's login and the end of that session. The edges' values represent the transition probabilities between two different states (or queries). Based on the Markov chain presented in Figure 3, we can see that, for example, the query S1 is the first query being made in 40% of the treated cases (this value is the label of the transition $S_0 \rightarrow S_1$) and queries S3 and S2 will be executed then, respectively, in 90% and 10% of all queries executed. The rule that support $S_1 \rightarrow S_3$ could be something like: $S_1 \rightarrow S_3$ *(sup= ...; conf=0.9)*.

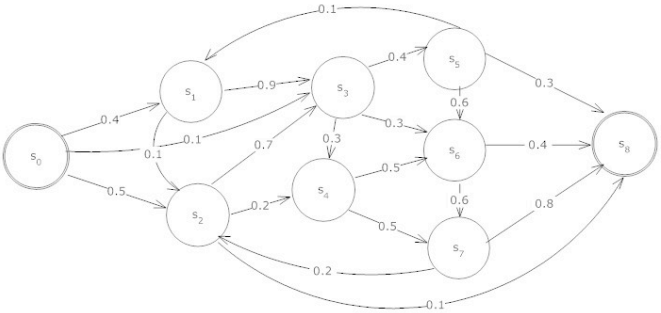


Figure 3. A query sequence prediction for the first dataset.

One way to reduce the number of queries to be materialized prior to the consultation (and cached) of a given user is by observing the probability of occurrence of all queries. Thus, from the start of a new working session, it is possible to identify which set of queries allows for reaching the final state (S8) with a greater probability of success. However, to do this, we have to look in each state of the Markov chain, which is the state most likely to be the next state to be reached until the final state is reached. All the tests conducted over this dataset basically used various values for *minconf* simplifying the rules accordingly. The chosen values for *minconf* were, respectively, 0.3, 0.4, and 0.5 (Table I). One other simplification was introduced, and named as "main route", simplistically put in cache the sequence of queries that a user will most likely follow in a future data exploration process, from login to logout.

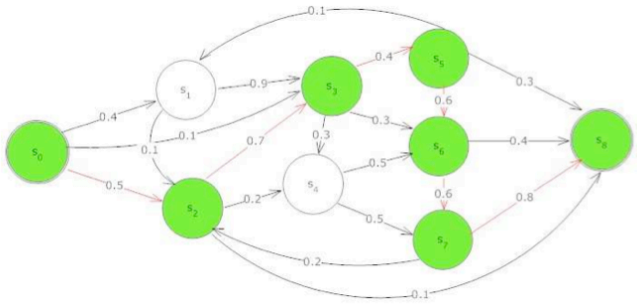


Figure 4. A candidate sequence of queries that an user most likely follow in a future data exploration process.

In Figure 4, we can clearly identify such candidate sequence of queries. It will be the sequence represented by the path:

$$S_0 \rightarrow S_2 \rightarrow S_3 \rightarrow S_5 \rightarrow S_6 \rightarrow S_7 \rightarrow S_8.$$

It can be easily found by following the higher transition probabilities between the S_0 and S_8 nodes. In the specific case of the Markov chain presented in Figure 4, the identification of the "main route" will be a result of the materialization of the nodes S2, S3, S5, S6, and S7, since nodes S0 and S8 correspond respectively to the begin and end session actions. The results of the tests, for the different values of *minconf* and for the "main route" simplification models, can be found in Figure 5. All the results of the tests were compared with each other - for a fairer comparison, we present the percentages for the values attained in each test.

TABLE I. TEST RESULTS FOR THE FIRST DATASET

Minconf	**0.3**	**0.4**	**0.5**	***"main route"***
Pre-materialized views (%)	100	86	28	71
Cache Hits (%)	100	89.8	38.3	79.78

As a comparison value, if we add 50% of all queries to the cache, intuitively we think we would achieve almost 50% cache hits for any given user (Figure 5). However, this value is merely meant to provide us with a reference value, and should not be considered in terms of absolute values. Figure 6 leads us to note two key values of *minconf* values if 0.3 and 0.5, which show the most relevant (best and worst) test results. As for the value 0.5, it means that only 28% of all possible views were pre-materialized and, even in that case, the cache hits came around 38.3%, which represents a 10% increase in system performance when compared to our reference values. The usage of 0.3 for *minconf* resulted in no view being simplified and, consequently, the values of cache hits were measured at 100%. In Figure 7 and Figure 8 we can see the simplified Markov chains that resulted, respectively, from the application of a *minconf* = 0.3, and a *minconf* = 0.5.

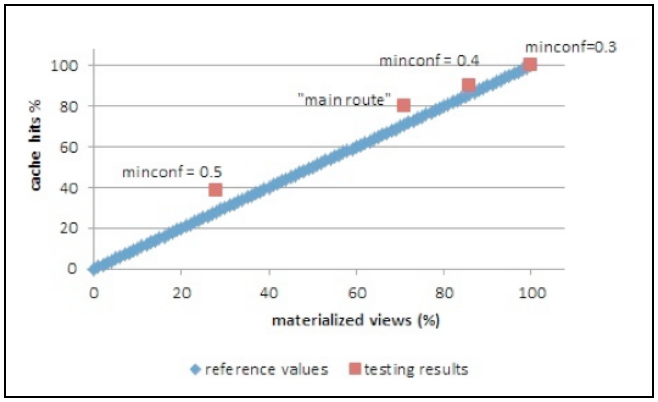
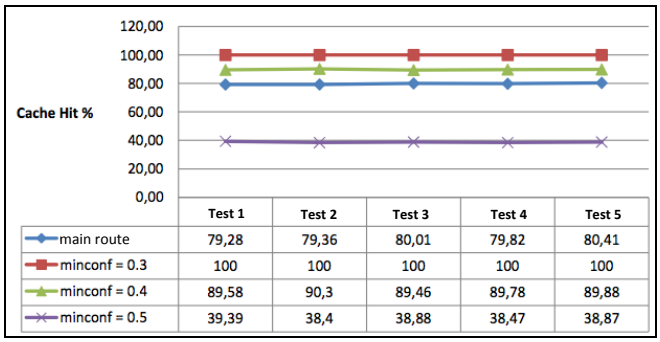


Figure 5. Test results graph for the first dataset.



	Test 1	Test 2	Test 3	Test 4	Test 5
main route	79,28	79,36	80,01	79,82	80,41
minconf = 0.3	100	100	100	100	100
minconf = 0.4	89,58	90,3	89,46	89,78	89,88
minconf = 0.5	39,39	38,4	38,88	38,47	38,87

Figure 6. Comparison of the results of the tests.

With the application of a confidence restriction of 0.3 it is possible to remove arcs that possess a lower probability than the value of confidence defined (Figure 7). Thus, it is possible to see that, for example, the arc corresponding to the transition between nodes S2 and S8 was not accepted. On the other hand, if the strategy is to materialize all views remaining in the Markov chain, then, with this confidence value, it is not possible to remove any of the views present in the chain. Due to this fact, it can be considered that although the definition of a more restrictive value of the minimum confidence (*minconf* = 0.3) the benefit in this case would be non-existent, since the set of views to materialize would be precisely the same.

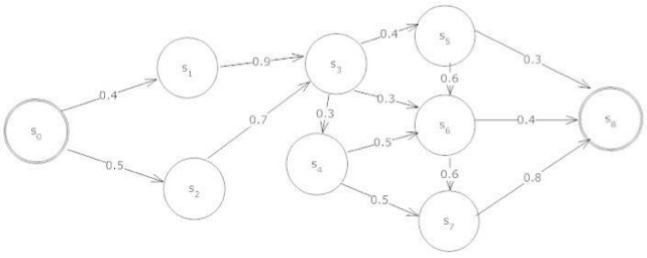


Figure 7. A simplified Markov chain for minconf = 0.3.

In the case presented in Figure 8 it is shown another simplification of the prediction model generated before, but now applying a *minconf* = 0.5. Following the logic

previously exposed for this case were removed from the model nodes S1, S4 and S5 since they did not have, after the first step of simplification, any incoming arc.

.

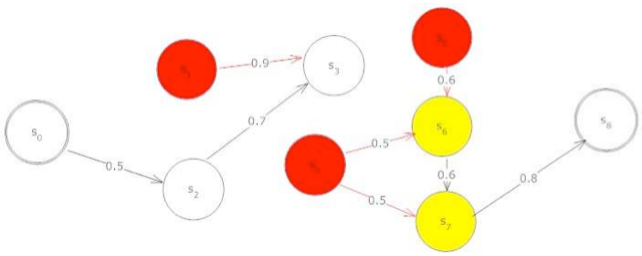


Figure 8. A simplified Markov chain for minconf = 0.5.

In a second stage of analysis, we observed that the node S6, after the removal of node S5, lost the only link it owned that could support the prediction of its occurrence. Because of this, it was necessary to remove as well the node S6. Then, with the same rationale, we removed the node S7.

TABLE II. TEST RESULTS FOR THE SECOND DATASET

Minconf	0.02	0.3	0.4	0.6
Pre-materialized views (%)	54	52	50	46
Cache Hits (%)	89	88	87	86

As can be seen from Figure 8 the restriction of a *minconf* = 0.5 has resulted in a quite considerable simplification of the prediction model previously generated. Thus, for this case, we will materialize only the views S0 (the initial state), S2, S3 and S8 (the final state). Later, other tests were conducted with another data set retrieved from several OLAP sessions we made on a specific OLAP server. This second dataset contains a total of 59 queries being issued to the server, and the values of *minconf* used to simplify the generated rules were 0.02, 0.03, 0.4, and 0.6. The results of this second experience can be found in Table II and Figure 9. The results obtained in this second round of tests shows us that, even though the differences between the different values of *minconf*, they do not yield great differences in the percentage of cache hits – nor in the percentage of materialized views. The gains relative to the reference values were quite relevant, staying approximately between 35% and 40% (for values of *minconf* equal to 0.02 and 0.6, respectively). Despite the lack of real caching data, all data sets prepared for testing and using in the several application scenarios we conceptualized provided the necessary means to prove the utility of our approach.

However, the results were not a surprise. In fact, based on the experience we have from other studies using association rules, we expected that the most frequent queries, as well as their more frequent sequences, were revealed naturally. This would allow for solid knowledge about analytical usage trends of a user community, and hence determine which query should be materialized in the cache at any given time. However, materialize only the queries indicated by association rules did not establish effectively a querying materialization plan for caching in the medium term. To support this in a more effective way we represented association rules in Markov chains. This allowed us to get a larger "horizon" for query materialization. Thus, this combination of techniques that provided us a very practical, not complex, way to establish *a priori* very practical querying materialization plans for an OLAP engine caching system, reducing consequently the workload of an OLAP server and improving its querying response time.

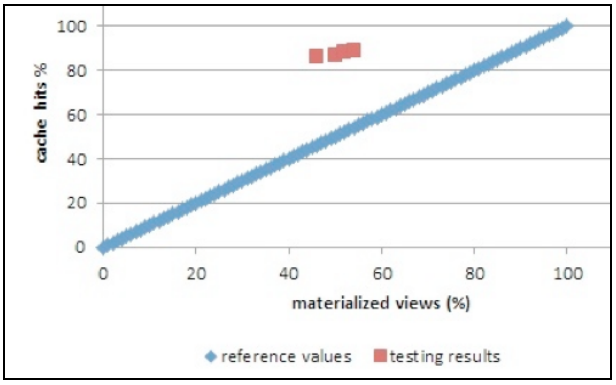


Figure 9. Test results graph for the second dataset.

Until the time we finished this work, we did not find any other caching solution for analytical systems such as the one we developed and exposed here. As such, we were unable to evaluate our proposal based on other alternative solutions that adopt or follow the same kind of strategy and categorization of patterns of analytic exploration we did. Even taking into consideration the works [14], [22], and [7], which revealed very curious and interesting aspects of predictive caching, we cannot make such comparative assessment.

VII. CONCLUSIONS AND FUTURE WORK

In recent years, the large increase using multidimensional database systems led to a greater interest in research of new techniques for the improvement of OLAP systems' functionalities and services. As widely used in other areas, these techniques can highlight the implementation and operation of caching mechanisms in OLAP systems, decreasing analytical server working load. As for Web systems, initially on OLAP systems the implementation of such mechanisms was only on the client side. This meant that only a certain user benefit of previous orders. Consequently, and as a manner to increase the advantage achieved by using caches, such mechanisms has focused on sharing benefits for all users of a given OLAP server. Thus, through the distribution of caches or by means of shared caches among individual users, either through the creation of systems specifically designed to serve as cache

servers, significant improvements were achieved in terms of performance [16][13][5][6][14].

The main goal of this study was to investigate in what conditions a predictive caching system could be used in a typical OLAP environment. In order to reach such goal, we studied several known cache techniques, e.g., [15][29][5][6] [25][12][14][7][4], trying to establish the basis to propose a different manner to know *a priori* the contents of an OLAP cache in a near future. All those techniques were crucial to the development of our work, for both the ideas of exploring the log files present in the OLAP Server and the simplification of the rules generated after the application of mining algorithms to that information. All the tests performed showed satisfactory improvements in the ratio between materialized views and cache hits. In our perspective, they also showed that this approach has the necessary pre-requisites to be applied to a more real scenario with advantages for the overall system's global performance.

The results of all tests demonstrated that the simplification of a certain percentage of rules to be pre-materialized does not mean that the same percentage of requests cannot be served from the cache. The doubt that remains is that if this number of cache hits is small in terms of a percentage higher or lower than the simplification of rules. Analyzing the data generated by tests, both using dummy data (the first data set), as the data retrieved from a specific data repository (the second data set), the technique for simplifying cache maintenance used proved to be very beneficial in all the tested scenarios. Even though some important questions remain, both for the period of logs that should be analysed and for the values of *minconf* to be used. This last, is an issue that should be addressed on a case-by-case approach, and should be included in a typical tuning-phase after finishing system implementation.

Finally, we think that with larger datasets feeding the mining algorithm, results should be even better. With a greater number of test cases, preferably from real application scenarios, it is possible to define with greater precision which are the access patterns of users as well as what benefits arise from the application of the several techniques. For that reason we plan in a near future to extend the current study, comparing it with other similar approaches and including some work concerning the exploration of multidimensional queries. We will give particular attention to the less busy periods of an OLAP server, in order to pre-materialize some specific multidimensional views that can be used latter when a user logs in – the log in periods can, as well, be subject of prediction. Yet because of the scarcity of the available data, and because it was decided not to integrate in this work the process where OLAP server logs are analysed and where are acquired all the association rules that serve as input to the technique developed – we focused on the simplification of the already generated rules –, we plan to, in a near future integrate these two phases in the process described in this paper. In the short term, we need to evaluate in a more effective way the practical utility of the predictive caching technique developed, extending the analysis periods along with the process of making a comparative study with other similar and concurrent techniques.

REFERENCES

[1] P. Marques and O. Belo, "Adaptive OLAP Caching, Towards a better quality of service in analytical systems," in Proceedings of The Second International Conference on Business Intelligence and Technology (BUSTECH'2012), pp. 42-47, Nice, France, July 22-27, 2012.

[2] A. Abelló and O. Romero, "On-Line Analytical Processing (OLAP)," in Encyclopedia of Database Systems (editors-in-chief: Tamer Ozsu & Ling Liu), Springer, pp. 1949-1954, 2009.

[3] N. Roussopoulos, Y. Kotidis, and M. Roussopoulos, "Cubetree: Organization of and Bulk Incremental Updates on the Data Cube," in proceedings of the 1997 ACM SIGMOD international conference on Management of data (SIGMOD '97), J. Peckman, S. Ram, and M. Franklin (Eds.). ACM, New York, NY, USA, pp. 89-99, 1997.

[4] W. Zhenyuan and H. Haiyan, "OLAP Technology and its Business Application, Intelligent Systems," in proceedings of the Second WRI Global Congress on Intelligent Systems, pp. 92-95, 16-17 Dec, Wuhan, 2010.

[5] P. Kalnis and D. Papadias, "Proxy-server architectures for OLAP," in proceedings of the 2001 ACM SIGMOD international conference on Management of data (SIGMOD '01), Timos Sellis (Ed.). ACM, New York, NY, USA, pp. 367-378, 2001.

[6] P. Kalnis, W.Ng, B. Ooi, D. Papadias, and K. Tan, "An adaptive peer-to-peer network for distributed caching of OLAP results," in Proceedings of the 2002 ACM SIGMOD international conference on Management of data (SIGMOD '02). ACM, New York, NY, USA, pp. 25-36, 2002.

[7] Q. Yao and A. An, "Using user access patterns for semantic query caching," in proceedings of Database and Expert Systems Applications, 14th International Conference. Prague, Czech Republic, 2003.

[8] M. Chrobak and J. Noga, "LRU is Better than FIFO," Algorithmica, Springer New York 23, pp. 180-185, 1999.

[9] V. Mookerjee and Y. Tan, "Analysis of a least recently used cache management policy for Web browsers", Operations Research, vol. 50, 2, pp. 345-357, Mar 2002.

[10] J. Boyar, M. Ehmsen, J. Kohrt, and K. Larsen, "A Theoretical Comparison of LRU and LRU-2," in Proceedings of the 4th International Workshop on Approximation and Online Algorithms, volume 4368 of Lecture Notes in Computer Science, pp. 95–107, Springer-Verlag, 2006.

[11] N. Megiddo and D. Modha, "Arc: a Self-Tuning, Lowoverhead Replacement Cache," in Proceedings of FAST '03: 2nd USENIX Conference on File and Storage Technologies San Francisco, CA, USA, March 31–April 2, 2003.

[12] W. Lehner, J. Albrecht, and W. Hümer, "Divide and Aggregate: caching multidimensional objects," in proceedings of the Second Intl. Workshop on Design and Management of Data Warehouses (DMDW 2000), Stockholm, Sweden, 2000.

[13] P. Cao, J. Zhang, and K. Beach, "Active Cache: Caching Dynamic Contents on the Web," Distributed Systems Engineering, vol. 6, 1, pp. 43-50, 1999.

[14] T. Loukopoulos, P. Kalnis, I. Ahmad, and D. Papadias, "Active Caching of On-Line-Analytical-Processing Queries in WWW Proxies," in Proceedings of the International Conference on Parallel Processing (ICPP '01). IEEE Computer Society, Washington, DC, USA, pp. 419-426, 2001.

[15] P. Deshpande, K. Ramasamy, A. Shukla, and J. Naughton, "Caching multidimensional queries using chunks," ACM.SIGMOD Rec. 27, 2, pp. 259-270, June, 1998.

[16] S. Bakiras, T. Loukopoulos, and I. Ahmad, "Dynamic Organization Schemes for Cooperative Proxy Caching," in IPDPS'03 (International Parallel and Distributed Processing Symposium), 2003.

[17] H. Gupta, "Selection of Views to Materialize in a Data Warehouse," in Proceedings of the 6th International Conference on Database Theory, Springer-Verlag, London, UK, 1997.

[18] V. Harinarayan, A. Rajaraman, and J. Ullman, "Implementing data cubes efficiently," ACM SIGMOD Record, vol. 25, 2, pp. 205-216, 1996.

[19] E. Baralis, S. Paraboschi, and E. Teniente, "Materialized Views Seleccion in a Multidimensional Database," in Proceedings of the 23rd International conference on Very Large Databases. Morgan Kaufmann Publishers Inc: San Francisco, CA,USA, 1997.

[20] A. Bauer and W. Lehner, "On solving the view selection problem," in Proceedings of the 15th International Conference on Scientific and Statistical Database Management. IEEE Computer Socienty, Washington DC, USA, 2003.

[21] Y. Kotidis and N. Roussopoulos, "A case for dynamic view management," in ACM Transactions on Database Systems. ACM: New York, USA, 2001.

[22] C. Sapia, "PROMISE: Predicting Query Behavior to Enable Predictive Caching Strategies for OLAP Systems," in Proceedings of the Second International Conference on Data warehousing and Knowledge Discovery (DAWAK 2000), Greewich, UK, Septeber 2000, Springer LNCS, 2000.

[23] K. Ramachandran, B. Shah, and V. Raghavan, "Dynamic pre-fetching of views based on user-access patterns in an OLAP system," in ACM SIGMOD, 2005.

[24] L. Cherkasova, "Improving WWW Proxies Performance with Greedy-Dual-Size-Frequency Caching Policy," HP Laboratories Technical Report HPL, 1998.

[25] M. Lawrence, F. Dehne, and A. Rau-Chaplin, "Implementing OLAP Query Fragment Aggregation and Recombination for the OLAP Enabled Grid," in proceedings of Parallel and Distributed Processing Symposium (IPDPS 2007), IEEE International, pp. 26-30 March 2007.

[26] J. Albrecht, A. Bauer, O. Deyerling, H. Günzel, W. Hümmer, W. Lehner, and L. Schlesinger, "Management of Multidimensional Aggregates for Efficient Online Analytical Processing," in Proceedings of the 1999 International Symposium on Database Engineering & Applications (IDEAS '99). IEEE Computer Society, Washington, DC, USA, pp. 156-164, 1999.

[27] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in J. Bocca, M. Jarke, and C. Zaniolo, editors, Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pp. 487-499, Santiago, Chile, September 1994.

[28] R. Howard, "Dynamic Programming and Markov Processes," MIT Press, June, 1960.

[29] S. Gribble, A. Halevy, Z. Ives, M. Rodrig, and D. Suciu, "What can databases do for Peer-to-Peer," in proceedings of WebDB, 2001.

Provide a Real-World Graph Suitable for the Mathematical Optimization of Communication Networks

Markus Prossegger
Department of Engineering and IT
Carinthia University of Applied Sciences
Klagenfurt, Austria
m.prossegger@cuas.at

***Abstract*—One of the main reasons of the hesitant expansion of fiber optic communication networks is the large financial expense for the excavation work. While the cost for (passive) hardware components and the fiber optics cable itself have fallen during the last decade, the cable layout work is still the cost driver of network construction projects. The most promising approach for a valid cost estimation is to use state-of-the-art network simulation and optimization techniques in the field of operations research. This study examines an automated process of generating a network graph that closes the missing link between the real-world and mathematical optimization of a communication network. The constructed graph is based on heterogeneous spatial data and weighted with real-world construction costs. It is then used to solve the minimum Steiner tree problem as typical optimization problem for the modeling and optimization of communication networks. While the applied mathematical model is studied in detail, the quality of the result as well as the runtime performance of the optimization algorithm is heavily dependent on the complexity and validity of the input graph. Based on a general format, the normalized geobasisdata, an initial graph, is constructed. This graph is then used as input into our rule-based system to select and weight the edges to be in the final graph. The optimization results of the experiments on real-world data, prove the effectiveness and efficiency of the proposed approach.**

***Keywords*—geographic information; spatial data; communication network optimization; normalized geobasisdata.**

I. Introduction

This study is an extension of [1] and examines the generation of network graphs, weighted with real-world construction costs to allow a valid optimization of communication networks. A typical network optimization problem is to connect a given number of customers by a wired network at a minimal expense. This state-of-the-art optimization problem is known as the Steiner tree problem (STP) that searches a cost minimal tree connecting terminals in a weighted network graph described in [2] and [3].

The generation of graphs, representing road networks, is subject of different approaches (e.g., the extraction of road networks from satellite images as described in [4] or the extraction of road intersections from raster maps described in [5]). The generation of real-world graphs, allowing the optimization of communication networks, is understudied.

To enable the mathematical simulation and optimization of a wired telecommunication network, the spatial data has to be converted into a weighted network graph, which is the missing link between the real world and the mathematical modeling. This graph is consisting of vertices (i.e., points-of-interest) and pair wise joining edges (representing network construction costs) between them. It is generated using spatial polygon data describing the land use on the one hand, and spatial line- and point-objects, describing existing infrastructure on the other hand. Based on this spatial data originating from a number of hybrid sources, a rule-based expert system is used to construct a network graph as vital input in subsequent mathematical models.

We focus on generating a weighted graph that can be used in different optimization models within the scope of wired telecommunication network construction. An instance of such a model using weighted graphs is the simulation and optimization engine of fiber optic communication networks described in [6]. Here, the land use polygons are divided into a grid, called cost raster. Each cell in this grid is representing the averaged underground construction costs determined by network constructors. While using a cost raster is a feasible way to generate a network graph representing network construction costs, a more sophisticated approach is needed to generate graphs by taking into consideration all kinds of real world information and being able to be computed in a reasonable time.

The present paper is divided into a preliminaries section, the section dedicated to the definition and origin of normalized geobasisdata, details and quality of our approach followed by experimental results and the conclusion.

II. Preliminaries

The subsequent simulation and optimization algorithms require undirected graphs as the fundamental data structure. The graph $G = (V, E, d)$ consists of $n = |V|$ vertices and $m = |E|$ edges. The distance of an edge $e_{ij} \in E$ connecting the two vertices $i \in V$ and $j \in V$ is given as a cost function $d_{ij} : E \to \mathbb{R}$, where $\mathbb{R}$ is the set of all real numbers. When calculating the shortest path or the minimum Steiner Tree as

typical routing problems, the distance d can be the Euclidean distance. The more sophisticated algorithms use travel time as the distance between two vertices. In case of scenarios considering the cable layout of wired networks, the distances have to be construction costs, such as underground work or the costs for building cable poles.

The goal of optimizing the cable layout of a communication network is to find a connected sub graph $S = (V_s, E_s)$ in G, connecting the terminals (i.e., access objects) $T \subseteq V$ such that the sum of edge weights $\sum_{e \in E_s} d_e$ is minimal. Verticals from the set $V \setminus T$ are called Steiner nodes.

The two-dimensional geographic data originate from hybrid sources, thus these data need to be prepared to serve as the basis for the construction of network graphs. There are three main sources for the geographical data:

(a) The geographic information system (GIS).
(b) The network information system (NIS).
(c) The digital cadastral map (DKM).

Each of the above listed items is needed for the construction of a consistent data source.

A. GIS

In our case, the geographic information system includes typical information used in marketing scenarios and strategic decisions. It is important to know where the potential customers are located. The data showing the population density or the number of households collected in a population census are incorporated in the GIS as well. The GIS contains statistical data aggregated from public sources together with information gathered by the prosecuting company itself. The most important information for a network construction company is the information about the location of potential, private or public customers as well as the expected benefit.

The network operator knows the exact location and the return on investment of the current customers, but not for potential customers. The marketing division uses market surveys and other statistical data to predict the location of potential customers and the likely yearly sales.

B. NIS

The network information system contains the information regarding all hardware components of the communication network as well as all logical links between these components. Typically, the NIS contains the most important business secrets of a network operating company. The following gives a list of the typical content of a NIS:

- Current and former customers.
- Network components.
- Physical cabling plan.
- Logical interconnection plan.

C. DKM

The Digital Cadastral Map is part of the official boundaries cadastre, which is the binding evidence of all parcel's boundaries. The DKM contains all public and private property and is typically available nationwide. Furthermore, it documents the type of land use of each parcel as well as buildings. Similar information is held in layers and together they form the DKM (a comprehensive interface description can be found in [7]):

- Boundaries.
- Parcel numbers.
- Types of land use (building land, forest, running water, standing water, etc.).
- Buildings.
- Control and Boundary points.

Formerly available only as an analog hard copy, the DKM was not only digitized but also enhanced using other official sources like Orthophotos and partition plans. Due to this reason, the quality of the digital map exceeds the quality of the analog version but it may include historical failures as well. There is a list of papers describing the aspects of spatial data quality [8], [9], [10], [11] as well as an ISO Standard regarding the quality of spatial data [12].

While the spatial accuracy is acceptable in most of mathematical simulation and optimization scenarios, the topological quality of the input data has to be ensured. There is some work proposed to identify spatial inconsistencies and incorrect object classifications using either manually defined spatial integrity constraints [13], [14], [15] or an automatic and incremental approach using decision trees proposed in [16] and improved in [17].

The DKM is used as one of the basic input into our approach and has to be normalized together with the other spatial input data.

III. Normalized Geobasisdata

The subsequent graph generation is designed as a completely automated process without the need of any user interaction. Due to this fact, the input data are stored in a predefined digital map format and the spatial objects must meet a set of conditions. In our approach, we have decided to use the ESRI Shapefile [18] as the digital map format. This open format is widely used and supported and stores spatial geometry and attributes as elements representing points, lines, and polygons.

The Normalized Geobasisdata (NGB) format [19] is an add-on to the ESRI Shapefile specifying the minimum qualitative and logical requirements of the spatial objects. It was developed in order to allow the automated generation of weighted network graphs based on any two-dimensional spatial data that represent surface data (i.e., land uses) in form of polygons at least. If the hybrid spatial data fulfill the specified NGB format, they qualify as input into the graph generation process.

The NGB format was originally developed for a simulation model dealing with the layout planning of a fiber-optics communication network in the year 2009. Since then it has

been continuously adapted to the specific requirements of individual projects.

The majority of mathematical simulation and optimization models dealing with cable infrastructure planning or routing in general rely on spatial data as the main input source to stay real world compatible. Table I represents the spatial information to be covered in the NGB format.

TABLE I
NGB OBJECT TYPES AND THEIR SPATIAL REPRESENTATION.

Id	Object class	Spatial representation
a	Project area	Polygon
b	Land use	Polygon
c	Usable (own or third-party) infrastructure	Polyline
d	Infrastructure points	Point
e	Access points	Point

The polygon describing the project area (a) as well as each polygon describing the land use (b) must show the following characteristics (in addition to some mandatory attributes described below):

- Valid and closed polygon.
- No crossing edges.
- Degree-two vertices only.
- No overlapping or equality with other polygons (see Figure 1).

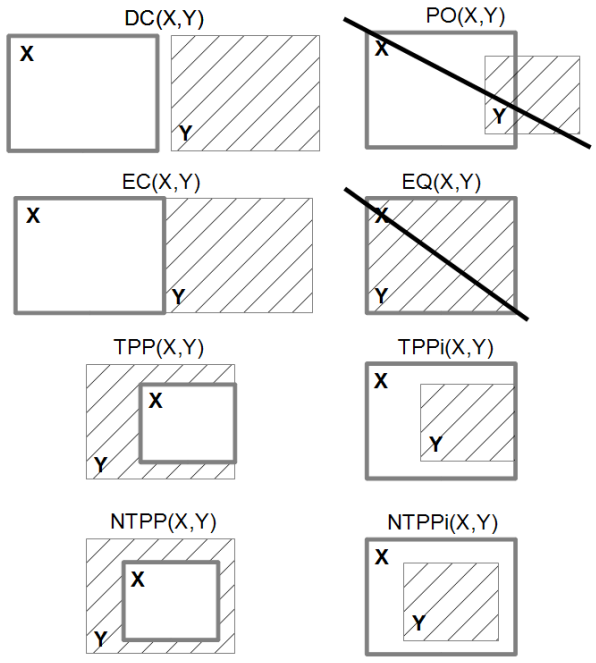


Fig. 1. All existing relations between pairwise polygons. EQ (equal) and PO (partly overlapping) relations are not allowed in NGB; Based on Region Connection Calculus (RCC-8) [20].

A polyline referred to as usable infrastructure (c) can be associated with own infrastructure (e.g., the copper cabling in the access domain) or third-party infrastructure (e.g., leased lines). The spatial object must be a valid polyline. In case of any attributive restriction in the accessibility, there must be at least two infrastructure points assigned; hence the infrastructure can only be accessed via one of the two points (an open line infrastructure can be accessed at the masts only).

The infrastructure points (d) are attributive assigned to exactly one polyline of the usable infrastructure and represent any type of infrastructure objects that can be localized on exactly one position (e.g., shafts to access pipes, masts, hardware like splitters or routers, etc.).

The last object class is the class of access points (e). These are points representing the terminals in a following optimization. We distinguish between existing access points that are currently supplied by one or a group of connected usable infrastructure polylines, and potential access points that are not yet connected.

The presented approach of the graph generation makes use of geometric algorithms and algorithms from operations research. Thus, ambiguous relations or even gaps between spatial objects cannot be allowed. This distinguishes between common geodatabase systems and spatial data in the NGB format, because there are no tolerances allowed.

Two points meant to be on the same position have to share the same coordinates. Furthermore, there are general specifications, which cover the notation, the coordinate system, the locale, default attribute values, and a list of common abbreviations.

The next section describes our graph generation approach, that is based on spatial input data in NGB format.

IV. GRAPH GENERATION

The process that we follow to generate a weighted network graph consists of three consecutive stages (see Figure 4 for the individual results):

- NGB preprocessing and enhancement.
- Generation of the candidate graph.
- Running the rule-based system.

In the following section, further details to the stages will be given.

A. NGB preprocessing and enhancement

The preprocessing and enhancement of the input data are fully automated processes. As long as the input data fulfill the requirements in the NGB format, the generation of a weighted network graph will succeed. Moreover, the quality and usability of the generated graph are crucial in terms of topological errors within the spatial data. Furthermore, the succeeding process of assigning the correct weights to all edges in the graph is sensitive to the correct spatial classification.

To ensure a valid real-world graph, the input data are validated running the decision tree approaches described in [16] and [17]. Based on error free spatial data covering provincial, rural, suburban, and urban areas, a representative decision tree was constructed. Both approaches use this decision tree to validate the input data. The process will output warnings in case

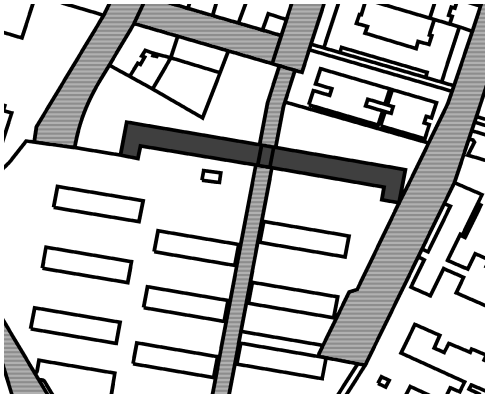


Fig. 2. The street (hatched area) is erroneously disconnected due to the building (dark filled area) that covers the thoroughfare.

of topological errors and reclassify spatial objects according to the decision tree. An example of an automatically identified error can be seen in Figure 2. The missing information about the thoroughfare can lead to a poor solution or even to an insoluble routing problem in consequence of the isolated street.

The validation is followed by the enhancement of the input data. Since underground work in crossroads areas and the subsequent obstruction in traffic should be avoided, the roadways are supplemented by polygons representing crossroads areas. Each crossing of at least two center lines of street polygons is identified and replaced by a polygon classified as crossroad area. The process runs automatically and produces results (see Figure 3), eminently suitable in the generation of valid candidate graphs.

B. Generation of the candidate graph

The goal of any graph-based mathematical optimization is to connect the access points to a given or new access network. As a result of the graph generation, each spatial object will be connected with other objects in the candidate graph. In the approach we follow three consecutive steps:

- Construction of the initial candidate graph.
- Classification of the graph edges.
- Assign real-world construction and activation costs.

Algorithm 1 describes the construction of the initial candidate graph that is used as an input into the rule-based system. Starting with the polygons representing the land use objects, the outer border and hole edges are collected and added to the graph. It is extended by the edges representing existing infrastructure considering restrictions in the accessibility. Additional projections originating from access points as well as artificial polygon crossings ensure a connected graph. The edges are classified with the corresponding land use and infrastructure running Algorithm 2.

Subsequently the classified edges are weighted with real-world construction and activation costs running Algorithm 3. The exemplary construction costs with respect to land use classifications and averaged activation costs of existing

(a)

(b)

Fig. 3. (a) Original NGB street polygons and (b) enhanced by crossroads area polygons.

communication infrastructure are shown in Tables II and III, respectively. The values are representing typical costs of underground work including the surface reconstruction, with the tendency to avoid airport and buildings as well as the crossing of railways, rivers, and lakes. Since the network constructing companies have their own empirical values, the construction and activation costs can be specified individually.

The rule-based system is applied to the generated and weighted candidate graph to significantly reduce the dimension (i.e., number of vertices and edges).

C. Running the rule-based system

The rule-based system is based on the expert knowledge of network constructors. It is applied to the generated candidate graph to test for qualified edges. Each of the following questions will be answered with yes or no and determine the appearance of the edge in the final graph (edge is rejected, if all answered with no):

Algorithm 1 : Generation of the candidate graph

```
Import all polygons P.
Import all infrastructure polylines L.
Import all infrastructure points I.
Import all access points A.
Import specifications regarding additional crossings.

for all polygons p ∈ P do
    Create edges representing the border of p.
    if p encloses an access point a ∈ A then
        Create (orthogonal) projections from a to p.
    end if
    if p should be enhanced with crossings then
        Create a crossing all x meter.
    end if
end for
for all polylines l ∈ L do
    if l has restricted access at two or more points i_{1..n} then
        Create edges between the points i_{1..n} representing l.
    else
        Create edges from the polyline l.
    end if
end for
for all infrastructure points i ∈ I do
    if i hits any created edge e then
        Split e into e_1 and e_2 at location i.
    else
        Create (orthogonal) projections to connect i.
    end if
end for
```

TABLE II
EXEMPLARY CONSTRUCTION COSTS WITH RESPECT TO LAND USE CLASSIFICATION.

Land use (extract)	Construction costs [€/meter]	
	along border	crossing
Agricultural land	50	100
Airport	1,000	1,000
Building	1,000	1,000
Building land	300	600
Freeway	140	500
Parkland	80	160
Plantation	200	400
Railway	100	1,000
River	100	2,000
Lake	250	2,000
Street	90	300
Vineyard	250	500
Woodland	400	400

1) The edge is needed to ensure a connected graph.
2) The edge is part of the existing infrastructure.
3) The edge is not part of any land use to be filtered (compare Figure 5).
4) The edge is part of a good (cost-efficient) possibility to connect spatial objects (access/infrastructure points).

TABLE III
EXEMPLARY ACTIVATION COSTS WITH RESPECT TO INFRASTRUCTURE CLASSIFICATION.

Infrastructure (extract)	Activation costs [€/meter]
Leased line	12
Own fiber optic cable (buried)	1,50
Own fiber optic cable (open line)	2

Algorithm 2 : Classification of edges

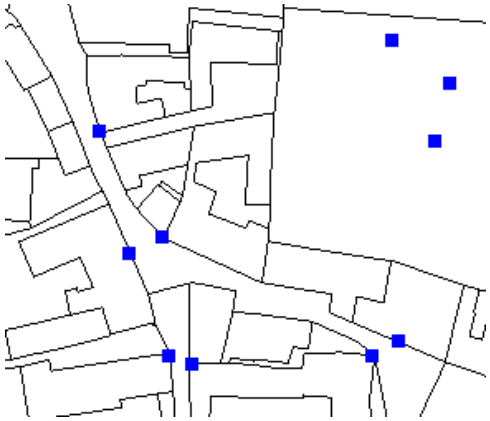
```
Import all edges E that are member in graph G.
Import all polygons P.
Import all infrastructure polylines L.

for all edges e_i ∈ E do
    if e_i intersects any other edge e_j ∈ E, where e_i ≠ e_j then
        Split e_i and e_j at the intersection point.
        Add splitted edges to set E.
        Remove e_i and e_j of set E.
    end if
end for
for all edges e ∈ E do
    if e is surrounded by polygon p ∈ P then
        Assign landuse(p) as classification to edge e.
        Label e as crossing edge.
    else if e is on border of polygon p_i ∈ P and p_j ∈ P then
        Assign landuse(p_i) and landuse(p_j) as classification to edge e.
        Label e as border edge.
    end if
    if e is part of infrastructure i ∈ I then
        Add type(i) as classification to edge e.
    end if
end for
```

Algorithm 3 : Weighting of edges

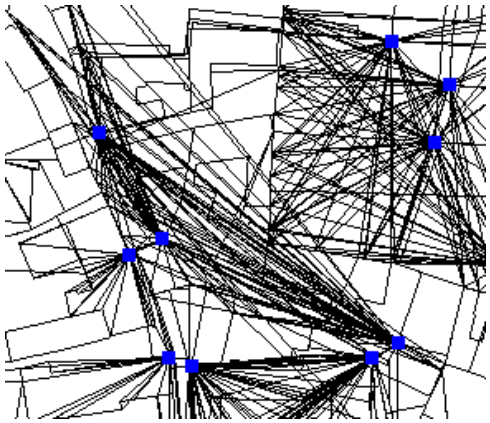
```
Import classified edges E that are member in graph G.
Import construction costs C_L from Table II.
Import activation costs C_I from Table III.

for all edges e ∈ E do
    Select the cheapest classification c of edge e considering the border/crossing label.
    Assign length(e) × c as weight to e.
end for
```

While for the first three questions the answer is quite easy to find, the last question requires running Algorithm 4 to be answered. The idea is to keep edges that allow the cost-efficient connection of spatial point objects regardless of the direction. Crossing edges will be discarded, if following the border is cheaper.

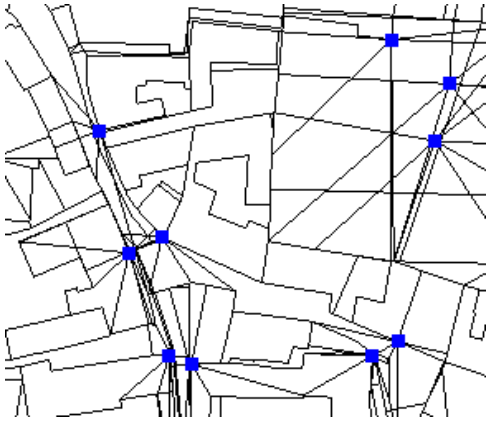
After identifying all edges that are members in the final graph, it can be used in the mathematical optimization of a



(a) Input data in NGB format



(b) Candidate graph



(c) Generated graph

Fig. 4. (a) Preprocessed and enhanced input data, (b) the candidate graph according to Algorithm 1, and (c) the resulting generated graph. Access points are shown as squares.

(a) Unfiltered graph

(b) Filtered graph

Fig. 5. (a) Unfiltered graph and (b) filtered by agricultural land uses.

Algorithm 4 : Cost-efficiency heuristics

```
Import weighted edges E that are member in graph G.
Import access points A.
Import infrastructure points I.
Import polygons P.
for all point objects o ∈ A ∪ I do
    Select all incident edges E_o ⊂ E of o.
    Cluster E_o into subsets representing 90° sectors.
    Keep the lowest weighted edge e ∈ E_o in each sector.
end for
for all polygon p ∈ P do
    Select all edges E_p ⊂ E, crossing p or following the
    border of p.
    for all edges e ∈ E_p connecting the vertices v_i and v_j
    do
        Discard e, if a shorter path between v_i and v_j is found
        in E_p.
    end for
end for
```

wired communication network.

V. Quality

The quality of the mathematical optimization result is heavily dependent on the complexity and validity of the generated graph. The main factors determining the quality of the generated network graph are:

(a) The usability.
(b) The real-world correlation.
(c) The correct cost assignment.

The (a) usability is primarily determined by the subsequent optimization algorithm. The dimension of the weighted graph (i.e., the number of vertices and edges) must be small enough to allow the application of heuristics or optimal algorithms but at the same time it also must be large enough to be non-restrictive. So a crossing of a river is a potentially unwanted scenario (due to the expense) but has to be made possible to reach isolated areas.

The (b) real-world correlation of the generated graph is equal to the correlation of the spatial data, if:

- all boundaries of parcels and buildings,
- all infrastructure, and
- all customers are contained and accessible.

To ensure the (c) correct weighting of the edges, the system is supplied with average network construction costs, broken down to activation/usage costs for infrastructure and excavation costs for each available land use. For the latter the granularity of excavation costs can be chosen freely, depending on the spatial information available. Due to the costs for a surface reconstruction or applying for official permits the crossing of a bitumenized street can be more expensive than to dig up a horizontal shaft along the boundary.

In the next section, we will describe the outcome of experiments we ran using real-world spatial data.

VI. Experimental Results

A series of experiments was run to evaluate the performance of the described graph generating approach as well as determining the quality of the weighted network graph. The algorithms were implemented using the object-oriented programming language C# (respectively the functional programming language F# implementing the rule-based system) from the .NET Framework version 4.5.1. All experiments were executed on a standard personal computer with Intel Core i5-2400 CPU, 8GB RAM, and 64-bit architecture.

A. Performance

To analyze the performance of the graph generation approach, we selected spatial data from four different classification areas:

- Urban;
- Suburban;
- Rural, and
- Provincial.

Table IV provides the classification of the exemplary selected input data together with the dimensions.

The dimension and runtime of the generated and weighted network graphs can be seen in Table V.

As expected, the runtime of the approach corresponds to the number of spatial objects enclosed in the area. For the simple reason that the graph needs to be generated only once, the duration is reasonable and acceptable.

TABLE IV
Selected classifications and dimension of enclosed objects.

Area	# Polygon Objects	# Line Objects	# Point Objects
urban	20,569	59,408	33,090
suburban	18,497	30,997	16,361
rural	5,255	21,850	11,800
provincial	792	1,076	508

TABLE V
Dimension and runtime of the generated graphs.

Area	# Edges	# Vertices	Runtime [sec]
urban	346^3	228^3	3,600
suburban	263^3	176^3	2,043
rural	139^3	88^3	882
provincial	17^3	14^3	519

B. Quality

The spatial input data selected for the quality experiments originates from the Austrian Digital Cadastral Map and covers a suburban area of about $6km^2$. It represents sparsely as well as densely populated areas to provide the most significant quality evaluation. To allow a self-sufficient comparison of the generated graphs, there is no existing infrastructure given. So the cabled communication network has to be build from scratch. The 42 points objects to be connected (i.e., Access points) are randomly selected centroids of buildings. Figure 6 shows the preprocessed and enhanced NGB of the area.

The generated graphs used in the experiments have been verified to contain all polygon boundaries (no filtering or aggregation of edges) and that there are no isolated parts. The weighting of the edges was done using average excavation costs per meter with respect to the underlying land use.

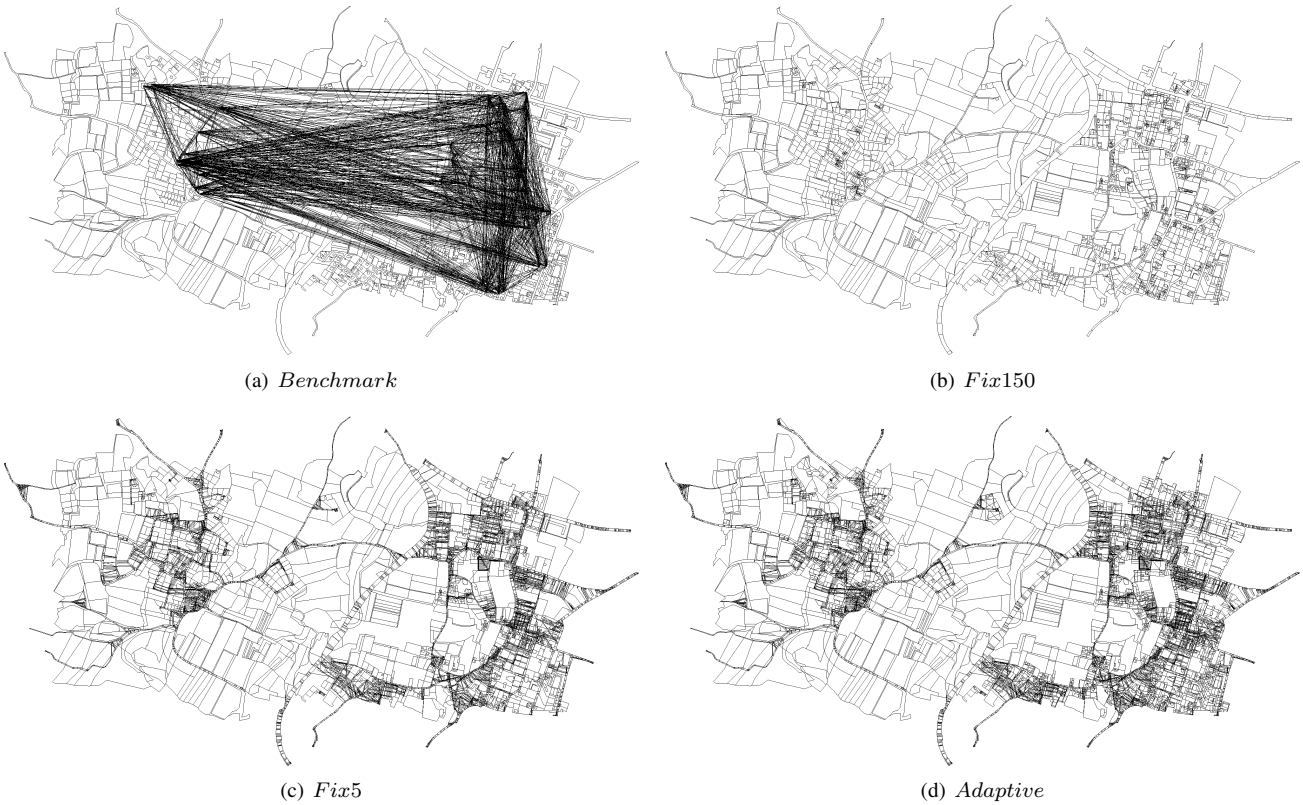
A fully connected graph (fully mesh all vertices) represents the best possible real-world correlation (because nearly every path is present) but the least quality with respect to the applicability. Considering only the polygon boundary, the number of vertices is about $13,300$ in our example. A fully connected graph would have 88^6 edges and therefore is not applicable.

We choose the best approximation as benchmark reference: a graph containing all polygon boundaries enhanced by edges building a fully connected subgraph of all access objects. Next to this we have generated three more graphs using our proposed approach.

The following list of network graphs were generated to allow an estimation of the quality:

- $Benchmark$ ... Polygon boundaries and fully connected access objects - Figure 7(a).
- $Fix150$ ... Generated graph with additional street crossings each $150m$ - Figure 7(b).
- $Fix5$ ... Generated graph with additional street crossings each $5m$ - Figure 7(c).
- $Adaptive$ ... Generated graph with adaptive crossings. The distance between two crossings is dependent on the construction costs of the underlying land use parcel.

Fig. 6. Spatial data of the project area representing the borders of land use polygons, street polygons (gray filled) and the access objects to be connected (black squares).

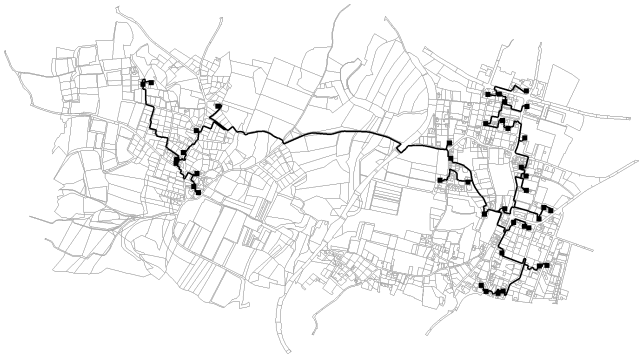


(a) $Benchmark$ (b) $Fix150$

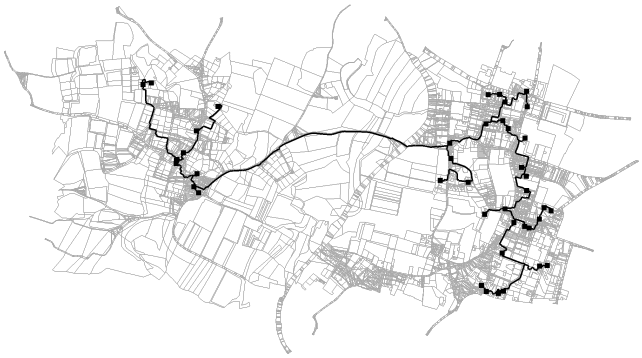
(c) $Fix5$ (d) $Adaptive$

Fig. 7. The graphs used in the experiments: (a) benchmark reference, generated graphs with fixed street crossings of (b) $150m$ and (c) $5m$, and (d) the graph with adaptive crossings.

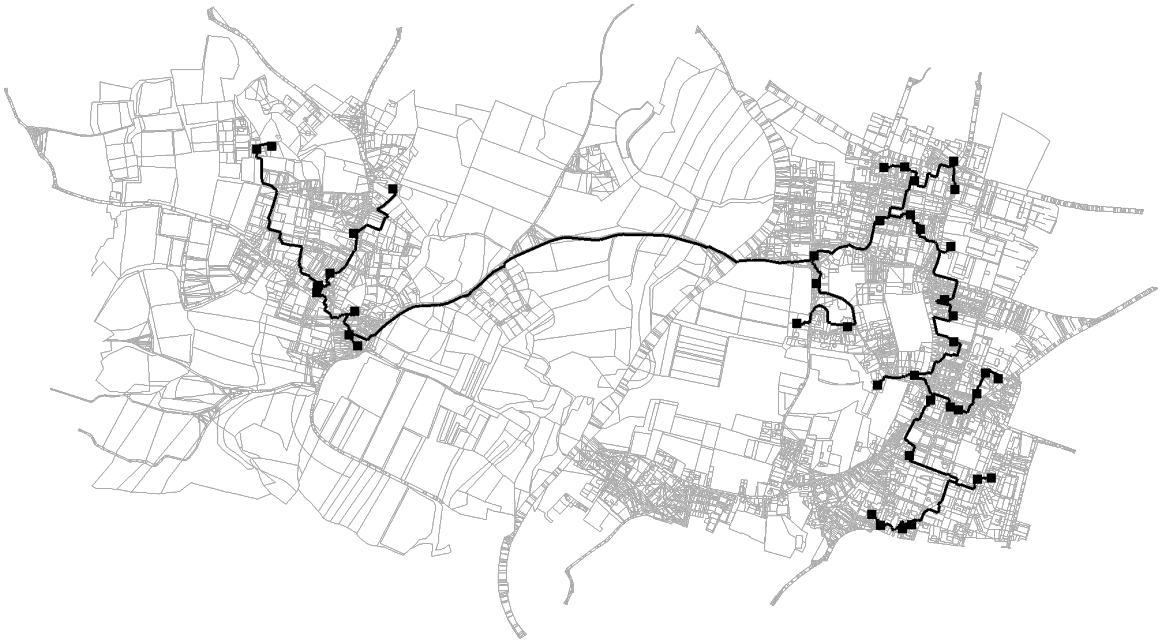
(a) $Benchmark$



(b) $Fix150$



(c) $Fix5$



(d) $Adaptive$

Fig. 8. The graphs superimposed by the optimized minimal Steiner tree running scenario (a).

Interpolated value between $10m$ (most expensive) and $500m$ - Figure 7(d).

Table VI is listing the dimension of the graphs. In case of the graph $Adaptive$, the number of edges and vertices is varying with the applied cost-set in the weighting process.

TABLE VI
DIMENSION OF THE GENERATED GRAPHS.

Graph	# Edges	# Vertices	# Access Objects
$Benchmark$	247^3	130^3	42
$Fix150$	17^3	13^3	42
$Fix5$	63^3	40^3	42
$Adaptive$	85^3 ($\pm 10^3$)	51^3 ($\pm 7^3$)	42

We used the defined graphs to approximate the minimal Steiner tree using the famous minimum spanning tree heuristics described in [21]. It has a worst case time complexity of $O(|T||V|^2)$ and guarantees a optimization result of no more than $2(1 - \frac{1}{l})$ times the result of the optimal Steiner tree (l represents the number of leaves in the optimal tree).

The following three different artificial scenarios (cost-triggered due to different cost-sets used in the weighting) have been applied:

(a) Each type of land use has its own cost value per meter. Edges on the boundary of a parcel and edges crossing a parcel show equal costs per meter.
(b) Avoid crossings of parcels. The costs per meter doubles for crossing edges.
(c) Force minimal usage of edges representing street crossings and avoid crossings of other land use parcels. The costs per meter doubles for crossing edges. The costs for edges crossing the street are multiplied by ten.

Table VII shows the optimized construction costs of a wired telecommunication network build from scratch. The costs of all edges in the Steiner tree are summarized in the third column. The last column is giving the deviation from the generated benchmark graph that is set as reference. In case of scenario (c) the deviation of graph $Fix5$ and $Adaptive$ is even negative. Hence, both the graphs are better suited for the optimization of a wired communication network than the complex benchmark graph.

Figure 8 is visualizing the optimized trees.

As shown in the experimental results, the proposed graph generation approach can be applied to generate weighted network graphs that are usable in mathematical optimization of network construction costs. The construction and activation costs are subject to change and should be updated prior to the edge weighting, to ensure a correct overall cost assignment. The real-world correlation is validated comparing the optimization results with a nearly fully connected graph.

TABLE VII
OPTIMIZED CONSTRUCTION COSTS

Scenario	Graph	Cost [Euro]	Deviation [%]
(a)	$Benchmark$	209,588	0
	$Fix150$	344,333	64.29
	$Fix5$	256,246	22.26
	$Adaptive$	242,198	15.56
(b)	$Benchmark$	306,502	0
	$Fix150$	426,692	39.21
	$Fix5$	346,584	13.08
	$Adaptive$	337,765	10.2
(c)	$Benchmark$	1,196,125	0
	$Fix150$	1,310,881	9.59
	$Fix5$	1,183,349	-1.07
	$Adaptive$	1,159,338	-3.08

VII. CONCLUSION

In this paper, an approach for the generation and weighting of a network graph has been proposed. Introducing a normalization format called NGB to support a fully automated process of generating graphs using a wide range of hybrid spatial data as the input. The process of generating a candidate graph followed by the classification and weighting of the edges produces a valid and computable real-world graph.

The experiments show that the approach is effective and efficient and that the weighted graph can be used as basic input into mathematical optimization algorithms. As a future investigation, we intend to explore ways to improve the adaptivity of the approach to further reduce the dimension of the generated graphs.

REFERENCES

[1] M. Prossegger, "Generation of a weighted network graph based-on hybrid spatial data," in *Proceedings GEOProcessing 2013, The Fifth International Conference on Advanced Geographic Information Systems, Applications, and Services*, 2013, pp. 120–124.
[2] R. D. Hwang, F. and P. Winter, *The Steiner Tree Problem*. North-Holland, 1992.
[3] A. Ivanov and A. Tuzhelin, *Minimal Networks: The Steiner Problem and its Generalizations*. CRC Press, 1994.
[4] O. Tuncer, "Fully automatic road network extraction from satellite images," in *Recent Advances in Space Technologies, 2007. RAST '07. 3rd International Conference on*, June 2007, pp. 708–714.
[5] Y.-Y. Chiang and C. A. Knoblock, "Automatic extraction of road intersection position, connectivity, and orientations from raster maps," in *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. GIS '08. New York, NY, USA: ACM, 2008, pp. 22:1–22:10.
[6] P. Bachhiesl, M. Prossegger, H. Stoegner, J. Werner, and G. Paulus, "Cost optimal implementation of fiber optic networks in the access net domain," in *International Conference on Computing, Communications and Control Technologies*, 2004, pp. 334–349.
[7] "Katastralmappe SHP Schnittstellenbeschreibung, Version 2.0.1." BEV - Bundesamt fuer Eich- und Vermessungswesen, 2012.
[8] R. Wang and S. D.M., "Beyond accuracy: What data quality means to data consumers," *Journal of Management Information Systems*, vol. 12, pp. 5–34, 1996.
[9] L. Leo Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Commun. ACM*, vol. 45, no. 4, pp. 211–218, 2002.
[10] B. K. Kahn, D. M. Strong, and R. Y. Wang, "Information quality benchmarks: product and service performance," *Commun. ACM*, vol. 45, no. 4, pp. 184–192, Apr. 2002.
[11] A. Jakobsson and F. Vauglin, "Status of data quality in european national mapping agencies," in *Proceedings of the 20th International Cartographic Conference*, vol. 4, 2001, pp. 2875–2883.
[12] "19113 Geographic Information - Quality Principles," in *ISO/TC 211*. International Organization for Standardization (ISO), 2002.

[13] T. Ubeda and M. J. Egenhofer, "Topological error correcting in gis," in *Proceedings of the 5th International Symposium on Advances in Spatial Databases*, ser. SSD '97. London, UK, UK: Springer-Verlag, 1997, pp. 283–297.

[14] K. A. V. Borges, C. A. Davis, Jr., and A. H. F. Laender, "Database integrity," J. H. Doorn and L. C. Rivero, Eds. Hershey, PA, USA: IGI Publishing, 2002, ch. Integrity constraints in spatial databases, pp. 144–171.

[15] M. Mostafavi, G. Edwards, and R. Jeansoulin, "An ontology-based method for quality assessment of spatial data bases," in *Proceedings fo the Third International Symposium on Spatial Data Quality*, vol. 28, 2004, pp. 49–66.

[16] M. Prossegger and A. Bouchachia, "Incremental identification of topological errors in spatial data," in *The 17th International Conference on Geoinformatics*, Aug. 2009, pp. 1–6.

[17] M. Prossegger and A. Bouchachia, "Incremental semi-automatic correction of misclassified spatial objects," in *Adaptive and Intelligent Systems*, ser. Lecture Notes in Computer Science, A. Bouchachia, Ed. Springer Berlin Heidelberg, 2011, vol. 6943, pp. 16–25.

[18] "ESRI Shapefile Technical Description," in *An ESRI White Paper*, July 1998.

[19] M. Prossegger, "Normalized Geobasisdata (NGB) - technical requirements v.3.0," FHplus Project Netquest, Carinthia University of Applied Sciences, Tech. Rep., October 2012.

[20] A. David, Z. Cui, and A. Cohn, "A spatial logic based on regions and connection," in *Proc. KR-92*, 1992, pp. 165–176.

[21] L. Kou, G. Markowsky, and L. Berman, "A fast algorithm for steiner trees," *Acta Informatica*, vol. 15, no. 2, pp. 141–145, 1981.

www.iariajournals.org

International Journal On Advances in Intelligent Systems
issn: 1942-2679

International Journal On Advances in Internet Technology
issn: 1942-2652

International Journal On Advances in Life Sciences
issn: 1942-2660

International Journal On Advances in Networks and Services
issn: 1942-2644

International Journal On Advances in Security
issn: 1942-2636

International Journal On Advances in Software
issn: 1942-2628

International Journal On Advances in Systems and Measurements
issn: 1942-261x

International Journal On Advances in Telecommunications
issn: 1942-2601