# International Journal on Advances in Intelligent Systems


2010 vol. 3 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 3, Number 3 & 4, 2010**

- Thomas C. Schmidt, University of Applied Sciences – Hamburg, Germany
- Karolj Skala, Rudjer Bokovic Institute - Zagreb, Croatia
- Chieh-yih Wan, Intel Corporation, USA
- Hoo Chong Wei, Motorola Inc, Malaysia

**Ubiquitous Systems and Technologies**

- Matthias Bohmer, Munster University of Applied Sciences, Germany
- Dominic Greenwood, Whitestein Technologies AG, Switzerland
- Arthur Herzog, Technische Universitat Darmstadt, Germany
- Reinhard Klemm, Avaya Labs Research-Basking Ridge, USA
- Vladimir Stantchev, Berlin Institute of Technology, Germany
- Said Tazi, LAAS-CNRS, Universite Toulouse 1, France

**Advanced Computing**

- Dumitru Dan Burdescu, University of Craiova, Romania
- Simon G. Fabri, University of Malta – Msida, Malta
- Matthieu Geist, Supelec / ArcelorMittal, France
- Jameleddine Hassine, Cisco Systems, Inc., Canada
- Sascha Opletal, Universitat Stuttgart, Germany
- Flavio Oquendo, European University of Brittany - UBS/VALORIA, France
- Meikel Poess, Oracle, USA
- Said Tazi, LAAS-CNRS, Universite de Toulouse / Universite Toulouse1, France
- Antonios Tsourdos, Cranfield University/Defence Academy of the United Kingdom, UK

**Centric Systems and Technologies**

- Razvan Andonie, Central Washington University - Ellensburg, USA / Transylvania University of Brasov, Romania
- Kong Cheng, Telcordia Research, USA
- Vitaly Klyuev, University of Aizu, Japan
- Josef Noll, ConnectedLife@UNIK / UiO- Kjeller, Norway
- Willy Picard, The Poznan University of Economics, Poland
- Roman Y. Shtykh, Waseda University, Japan
- Weilian Su, Naval Postgraduate School - Monterey, USA

**GeoInformation and Web Services**

- Christophe Claramunt, Naval Academy Research Institute, France
- Wu Chou, Avaya Labs Fellow, AVAYA, USA
- Suzana Dragicevic, Simon Fraser University, Canada
- Dumitru Roman, Semantic Technology Institute Innsbruck, Austria
- Emmanuel Stefanakis, Harokopio University, Greece

**Semantic Processing**

- Marsal Gavalda, Nexidia Inc.-Atlanta, USA & CUIMPB-Barcelona, Spain
- Christian F. Hempelmann, RiverGlass Inc. - Champaign & Purdue University - West Lafayette, USA
- Josef Noll, ConnectedLife@UNIK / UiO- Kjeller, Norway
- Massimo Paolucci, DOCOMO Communications Laboratories Europe GmbH – Munich, Germany
- Tassilo Pellegrini, Semantic Web Company, Austria
- Antonio Maria Rinaldi, Universita di Napoli Federico II - Napoli Italy
- Dumitru Roman, University of Innsbruck, Austria
- Umberto Straccia, ISTI – CNR, Italy
- Rene Witte, Concordia University, Canada
- Peter Yeh, Accenture Technology Labs, USA
- Filip Zavoral, Charles University in Prague, Czech Republic

**International Journal on Advances in Intelligent Systems**
**Volume 3, Numbers 3 & 4, 2010**

**CONTENTS**

# Game-Based 3D Simulation of Life in the Middle Ages for the Edutainment in Cultural Heritage

The reconstruction of medieval Otranto

Lucio T. De Paolis, Giovanni Aloisio
Department of Innovation Engineering
Salento University
Via Monteroni
Lecce, Italy
lucio.depaolis@unisalento.it
giovanni.aloisio@unisalento.it

Maria G. Celentano, Luigi Oliva, Pietro Vecchio
Scuola Superiore ISUFI
Salento University
Via Monteroni
Lecce, Italy
mariagrazia.celentano@unisalento.it
luigi.oliva@isufi.unile.it
pietro.vecchio@unisalento.it

*Abstract*—**Virtual Reality applications on Cultural Heritage are increasing, according to a general trend towards virtual reproduction and interaction mediated by the computer system. The effects of this trend, both on education and research, are still far from being completely tested and defined. The aim of the MediaEvo Project is to develop a multi-channel and multi-sensory platform for the edutainment in Cultural Heritage, towards integration of human sciences and new data processing technologies, for the realization of a digital didactic game oriented to the knowledge of medieval history and society. The developing of the project has enhanced interactions among historical, pedagogical and ICT researches, morphological inquiries, data management systems, by means of the definition of a virtual immersive platform for playing and educating. The platform is also intended to collect feedback and validate hypothesis and findings coming from researchers. This essay introduces the questions related to the educative use of ICT and describes the steps of the reconstruction of the town of Otranto in the Middle Ages: data collection and integration, organization of work and software applications.**

***Keywords- Simulation; Edutainment; Virtual Cultural Heritage; Urban History***

## I. INTRODUCTION

There is a worldwide interest in for Virtual Reality (VR) technology in Cultural Heritage in order to recreate historical sites and events for such purposes as education, special project commissions and showcase features at visitor centers.

The power of VR lies in its ability to open up places and to see things not normally accessible to people. VR also allows users to explore objects and to experience events that could not normally be explored without alterations of scale or time. The user can actively participate in creating new knowledge by doing and interacting with other users and objects in the virtual environment.

The use of VR has defined new fields inside traditional research contexts. Today we consider virtual archaeology, virtual architecture or urbanism and so on as defined disciplines specialized in enhanced virtual representation or reconstruction as a distinctive methodology of approach.

One of the best uses of virtual models is creating an environment to help students to learn about ancient cultures and to interact in a new way, using many possibilities for collaboration, in a shared social space.

Recreating or simulating ancient cultures, virtual heritage applications create a bridge between historic characters and contemporary users.

There are many experiences of historical environment reconstruction, the most successful are available on the web or have been presented in international conferences. Some of them relate to the elaboration of models or algorithms for better representing and reconstructing important sites, others explore Augmented Reality applications for Cultural Heritage, others test ontological systems for data managing and sharing**.**

It is a widely held point of view that cultural heritage is diminishing continuously. While new treasures emerge from places previously unexplored or ignored, a larger number of buildings and sites are compromised by natural or human action. This process leads to the demise of important historical documents and artistic goods.

The improvement of technological capabilities enriches the possibilities for research and protection and enhances the value of cultural heritage, thus halting their demise.

Firstly, the increased speed of communication, data exchange and data processing offers to the research community the dimension of real-time interconnectivity.

Secondly, the overall amount of information originating from both qualitative and quantitative exploration with the support of technologically advanced equipments, compared with that of a few decades ago, leads to the possibility of an extremely detailed description of reality.

The systems for cataloguing and managing these data have been structured with complex and ontological categories (the term *ontology* refers to a "specification of a representational vocabulary for a shared domain of discourse - definitions of classes, relations, functions, and

other object" [1]) that define common protocols for enhancing classification and comparison, even among distant users.

Finally, the elements that constitute the overall sign of the times are the possibilities presented by the means for a realistic representation of everything that comes from research, from the hyper sensorial reproduction of reality to the reconstruction of different hypotheses and scenarios.

The expansion of these means necessitates a contextual disciplinary revision of interest to all those in the field of humanistic studies. Historians cannot afford to buck the trend to a post-literary dimension of knowledge transmission or knowledge itself [2].

The new phase of contemporary civilization has been defined post-modern or, more correctly, post-historic [3], for the predominance of representation and hard virtualization of reality.

Evolution in research methodology corresponds to a general debate on communication and education closely linked to the characteristics of a changing perception of teaching, oscillating between experimental impulses and conservative attitudes [4], [5], [6].

The approach to the historical city, in terms of research and understanding both academic and popular, has been enriched by new tools and thanks to the development and proliferation of advanced technologies. The speed with which the use of computers and electronic devices has grown by a very wide range of users demands constant progress in the ways in which information is gathered, managed and transmitted.

On one hand the virtualization of space has reached such levels of mimesis as to influence the perceptual field and the capacity for evaluation of the experience. On the other hand, this factor, which is destabilizing for the whole field of investigation and understanding of reality – already predicted nearly half a century ago by various authors –, has completely changed the praxis and market for entertainment. Thus positively influencing the rate of contextual assimilation of the information but also negatively affecting concentration spans for its reception.

The new post-historical and communicative research frontier that interprets these processes focuses on the possibility of increasing a multidisciplinary approach and interchange, reacting in an active way to the promotion of cultural heritage and safeguarding against the degenerative processes that undermine it.

Communication and transfer of data nowadays occurs in real time, making an infinite amount of information available originating from quantitative and qualitative investigation. New systems for cataloguing and managing this data are organised according to structures of reference which are of a formal-ontological nature and ever more complex for which common protocols evolve facilitating the identification and comparison even of realities which are quite distinct from each other.

Finally, the already immense possibilities for realistic representation of all that emerges from research are growing, from the information that is detected with extreme precision to the different reconstructive hypotheses.

On this basis, the MediaEvo Project is aimed at creating a multichannel and multisensory edutainment platform for Cultural Heritage, through the integration of the human sciences and Information and Communication Technologies (ICT).

The activities include the creation of an educational video game aimed at spreading knowledge of medieval society in the area of Salento through the reconstruction of the city of Otranto in the XIII century.

## II. RECONSTRUCTING HISTORICAL CONTEXTS

The virtual reconstruction of a historical landscape can be divided in five levels [7].

*1)* The first level, *Archaeological Landscape*, regards all the information coming from physical measurement (we choose to call it, properly, *Realscape*).

*2)* The second level is the *Interpreted Landscape* or *Mapscape* that is defined by the systematic organization of data.

*3)* The third one is the hypothesis of a possible landscape in the past, *Ancient Potential Landscape* or *Pastscape*.

*4)* The fourth level involves the experience of historical context through a process of immersion which defines contemporary perceptions. With the aid of the social sciences this leads to the definition of Perceived Landscape or Mindscape.

*5)* The final level is the Webscape, the grid of outer relations and communication that is useful to test the process and collect the necessary feedback.

In the academic world, the "historic vision" is usually limited to professors, scholars and researchers, who share the interpretation codes for extracting the ancient landscape from the actual one. In this new stream of experimentation, geared towards interaction and edutainment, the researcher finally becomes part of a system through which to study and interpret space.

In a virtual interactive town, the possibilities of information exchange increase dramatically from the static reconstruction to the simulation. The simulation allows the construction of a platform that adds the definition of game rules and plots to interaction and immersion. This allows players to easily experience and recognize topographical and temporal coordinates of virtual space. In this way the past is actualized with real behaviors, producing at the same time, the vision of pastscape and mindscape in the virtual reality built on realscape and related to the mapscape.

Studying a town and its historic landscape involves different methods of analysis, interpretation and communication using digital technologies. Geographical Information System (GIS), remote sensing, laser scanning, photogrammetry, computer vision, 3D modeling, Virtual Reality (VR) and Augmented Reality (AR) are instruments

of a multidisciplinary system that links historical knowledge, structural recognition, geotopography, geology, sociology, urban and architectonic analysis, engineering and graphic skills.

The ancient town, as an information unit, can be defined as a *meme* [8], a cultural unit code that locates and describes the process of territorialization of human society. It is the space-time relation between man and environment at a certain time [9].

## III. RELATED WORK

The methodological and disciplinary peculiarities concerning VR have opened up new possibilities within disciplines that have led on, in the space of only a few years, to develop distinct characters of their own. Now days, we speak of virtual archaeology, virtual architecture, virtual town planning and so on, indicating that part of the discipline which is closely linked to material contexts and specialized in the reconstruction or verification of classical assumptions or of new hypotheses. The pure humanistic disciplines (history, philosophy, etc.) are still some way from this point. Their contribution, however, is fundamental in order to validate all the work in this environment.

Several VR applications in Cultural Heritage have been developed, but only very few of these with an edutainment aim.

Song et al. [10] present the historical and cultural content of the reconstructed 3D VE to the general public in a pedagogical and entertaining way; they incorporate interactive storytelling techniques into a Digital Heritage application. Because they believe interactive storytelling techniques can enrich the process of exploring the VE since each visitor can walk away with a different virtual experience.

Kiefer et al. [11] describe a subclass of location-based games, Geogames, which are characterized by a specific spatial-temporal structuring of the game events and assert that spatial-temporal structuring makes it easy to integrate educational content into the course of the game.

Cutrì et al. [12] study the use of mobile technologies equipped with global positioning systems as an information aid for archaeological visits. They conclude that the use of this kind of technology is an effective tool to promote the archeo-geographical value of the site.

Luyten et al. [13] present Archie, a mobile guide system that uses a social-constructionist approach to enhance the learning experience for museum visitors. They created a collaborative game for youngsters that is built on top of a generic mobile guide framework. The framework offers a set of services such as a rich interactive presentation, communication facilities among visitors and the possibility to personalize the interface according to the user group.

In the work of reconstructing historical or archaeological landscapes, extensive experimentation takes place on the net or has been presented during the course of international conferences. These primarily concern the elaboration of algorithmic models in order to better comprehend and reconstruct the sites, technological applications for AR applied to cultural heritage and ontological systems and data management.

An example of activity in the fields closest to the object of the present research is the work of the Institute for Architectural and Monumental Heritage [14] which has produced various reconstructions of the city of Metaponto (in the province of Matera) and of Muro Leccese (in the province of Lecce) and the monasteries of Santa Maria of Cerrate (province of Lecce) and Jure Vetere in the medieval age [15].

The reconstruction of the site of Faragola (province of Foggia) by the University of Foggia, undertaken as part of the Project Itinera [16]and known as *Time Machine*, fits within the trend of an experiential relationship within an archaeological context.

Other applications facilitate access to and reading of the cultural patrimony both within the museum and online: *Appia Antica Project* [17], *Virtual Rome Project* [18], *Muvi,* a virtual museum dealing with daily life in the XX century [19], and *Nu.M.E. Project*, a virtual museum concerned with the city of Bologna [20], are all to be considered prominent examples of experience relating to the latter.

On a strongly interactive level and related specifically to multichannel edutainment, examples of applications utilizing Virtual Collaborative Environments (CVEs) are found in the platform *City Cluster* [21] that permits the user to share in a virtual visit of various cities, *Quest Atlantis Project* [22] for teaching about archaeological contexts and *Integrated Technologies of Robotics and Virtual Environment in Archaeology Project* [23] that indicates a more professional use of VR interaction aimed not only at information dissemination but also scientific examination.

## IV. THE MEDIAEVO PROJECT

The MediaEvo Project aims to develop a multi-channel and multi-sensorial platform in Cultural Heritage and to test new data processing technologies for the realization of a digital didactic game oriented to the knowledge of medieval history and society.

The game is intended as a means to experience a loyal representation of the possible scenarios (environments, characters and social roles) in the historic-geographical context of Otranto during Swabian Age (XIII century).

We chose Otranto as an example town; Otranto is located in the south of Italy. Due to its geographical position (in the extreme East of Italy), Otranto was like a bridge between East and West.

The implementation of the edutainment platform is strongly influenced by the definition of the scenery that is the world in which the framework is placed with the related learning objects and learning path, the characters, the scene's objects, the logic and so the rules of the game, the audio content, the texts and anything related to its use.

The framework will have features of strategy games, in which the decision-making capacities of a user have a big

impact on the result, which in our case is the achievement of a learning target. Nevertheless the strategy and tactics are in general opposed by unforeseeable factors (provided by the game) connected with the edutainment modules, in order to provide a higher level of participation, which is expressed in terms of the ease with which it is learnt. The idea is to provide a competition between the players, during the learning.

The system, on the basis of a well defined learning target and eventually based on knowledge of the user, will continuously propose a learning path (learning path composed of a sequence of learning objects), in order to allow the achievement of particular learning results.

## V. MEDIEVAL OTRANTO AS A SCENERY

The city of Otranto was identified as a unique and eloquent historical setting for the project. Although the specific field of research was focused on the late middle ages, the project is set in a site which has been densely inhabited since before the VII century BCE and which conserves the signs of the previous cultural stratifications.

In figure 1 is shown a bird's eye perspective of the old town of Otranto [24].

The project leaves open the possibility for further work on other historical phases with the prospect of developing a complete ‘time machine’.

Through its art, spatial relationships and landscape, Otranto provides evidence of the close contact between Mediterranean cultures, particularly those of western Roman Catholicism, Byzantium and Islam. The year considered representative for the medieval reconstruction is 1227 - the year in which Emperor Frederick II of Swabia and his court entered the city for the first time to embark for the Sixth Crusade.

From the analysis of the monuments and documents, numerous useful points that facilitate the multicultural experience emerge to enrich the educative platform of immediate reference.

Otranto was officially a bilingual city. Together with Latin, Byzantine Greek was officially spoken by the archbishop during religious celebrations and both languages were taught at San Nicola of Casole - one of the great centers of cultural conservation and diffusion known as the *scriptoria*.

Being a maritime and mercantile city, the languages of the populace were many and varied. Throughout its history, Otranto has been settled by cultures that have influenced on the city on both an historical and artistic level [25], [26].

This cultural melting pot produced a particular mix of knowledge and traditions, still recognizable in some of the customs, handcrafts, and figurative art and in the articulation of space in Otranto. Interaction in a local context of this kind cannot but represent situations that resonate with the great themes of medieval civilization, in a sort of tiny virtual encyclopedia.

Figure 1. Bird's eye perspective of the old tonw of Otranto.

## VI. THE STEPS OF RECONSTRUCTION

The ancient town of Otranto preserves relevant elements that witness Middle Ages culture but also the former and latter ones. This could increase the pedagogical purposes and place the project in a more complex, complete, "time machine" perspective.

In 1227, when the Emperor Frederick II of Swabia, entered the town with his wife and court, Otranto was a cultural melting pot. Even if there were two official languages, latin and byzantine greek, walking in the town it wasn't uncommon to hear people talking in hebrew, armenian, vareg, french, provencal, german, arabic, etc.

All those elements can be reflected in a big deal of situations that are useful for educational purposes. In other terms we could say that Otranto, as it is represented in the game, becomes a compact, interactive, little encyclopedia of Middle Ages civilization.

### *A. Data acquisition*

The general information we first collect are actual Digital Terrain Models (DTM), thematic, technical, hydro-geological, nautical charts. On local side, surveys and metering operations produced maps of street organization, urban limits and fortifications, monuments and materials, referenced to absolute coordinates (*mapscape*).

Information coming from archaeology (published or available in archives) has been inserted in topographic charts, distinguishing the different historical period [27], [28].

The overall amount of data acquired and represented defines the actual state and conformation of the town (*realscape*) on which we are making a process of subtraction (reverse stratigraphy), in order to obtain the urban fabric on year 1227.

Unfortunately, during the last centuries, there has been a substantial loss of historical documents. The survived ones are not enough to describe efficiently the town in Middle Ages.

The first views of Otranto date to the end of XV century. They are more symbolic than realistic. Furthermore, historical maps and views have been collected and classified, together with relevant documents and plans.

### B. Data interpretation

The numerous gaps regarding, above all, the urban structure and placement of notable building, monumental and functional contexts were filled in part by a historical-urban and architectonic analysis in order to establish the spatial hierarchy, the urban poles, the lot sizes and the typological distribution [29].

The material elements were compared with analogous situations relating to surrounding areas or cities and modulation grids on a typological-functional basis were used for the built environment, the objects, the clothes and activities.

The possible scenarios for the era in question were added to the base consisting of the above-mentioned data (*pastscape*). This is updated in real time, little by little as the extent and detail of the research and representation is extended and enriched.

### C. The creation of the urban landscape

The first phase of the reconstruction involved the use of GIS in order to model the georeferenced DTM, on the basis of the reconstruction of the hypothesized altitude and sea level of the time. On this, the extra-urban roadways were identified, defining the hierarchy of pathways and their structural characteristics (stone, pressed earth, rock, etc) and relating them to the presumed location of the port.

Reasoning on the basis of vicinity and typology, the settlement maps for the various homogenous parts of the city were made, starting with the area around the Saint Peter's church.

The architectonic elements reconstructed were made using two modalities. The existing monumental buildings were modelled on the basis of a critical reconstructive survey which rendered them in their XIII century state, with what is supposed to have been lost at the hand of degrade, maintenance or restoration integrated into the reconstruction.

The curtain walls and the residential units of different types were based on an analysis of the city and the metrics of the time [30], they were reproduced, catalogued and entered into a database. For every one of these a set of variations was foreseen (form, composition of levels, openings, mouldings, surfaces, materials, etc) to distinguish them and promote a realistic perception of the game.

The historic scenario is however a static representation of a context. The final goal of interactive reconstruction is the definition of an immersive platform able to let players experience and feel the socio-cultural values of that period (mindscape). This is reached towards the creation of high representative interactive contexts:

- defensive (fortifications, castle);
- religious (the diocesan space: cathedral-tower bell-square and the churches);
- infrastructural (function and hierarchy of road axes, identification of the central distribution system and its links);
- commercial (buildings and structure devoted to exchange, commerce, distribution and collection of goods);
- intermodal (port, regional roads);
- sub-urban (expansion areas, non urban functions: monasteries, docks, fields, etc.);
- residential (neighborhood, social-economical-racial concentration and building types);
- artisan (arts and crafts);
- familiar.

### D. Analysis and findings

#### 1) The walls

The defensive context, which includes the city walls, the castle, the internal and external garrisons for surveillance and responding to attacks, is to be considered the first context of reference for reasons of both a cultural and spatial nature.

By definition in contemporary historiography: a city is an urban area surrounded by a ring of walls, inside which men of different families and occupations live without interruption [31].

Until the present day, the city has preserved its defensive structure that is the result of sudden, radical defensive reorganization after the Christian recapture of the city after their tragic expulsion by the Ottomans in 1480.

The age and the impact of this intervention don't allow us to effectively determine the original medieval image. A more circumstantial investigation is required. In relation to the historical documentation, the findings of archaeological campaigns and above all on the basis of the analysis of the urban fabric, some useful considerations relevant to the reconstruction can be expressed.

The historical centre of Otranto is located on a strip of land between two watercourses. The natural elevation rises to an altitude of approximately 35metres above average sea level (in the area of the current cemetery) and falls to an average of 14 metres in the ancient city, fronting onto the sea at 12 metres.

The coastline near the city centre is characterised by an inlet, which corresponds with the two outlets for the water channels and results in a double internal cove that the promontory of land constituting the residential centre overlooks.

Between these, the larger of the two water basins, the Idro, is a fundamental element for the entire settlement (he probably gave the name to the town, according to someone) for the fact that it guarantees a minimal, continual supply of water in a region which is characterized by an intrinsic lack of surface water.

In figure 2 is shown the definition of hydrographical, urban and defensive structures of the Otranto town.

The geomorphology and hydrology of the site identify a natural system that contains within it and influences the characteristics of the residential centre both on a functional level, in terms of the infrastructure and – above all for the ancient and medieval eras – on a strategic-defensive level.

The archaeological evidence reveals a substantial continuous settlement located on the shore of the sea, which dates to the Messapian era, made evident by the surviving fragments of city walls – often reinforced – brought to light in the course of archaeological excavations.

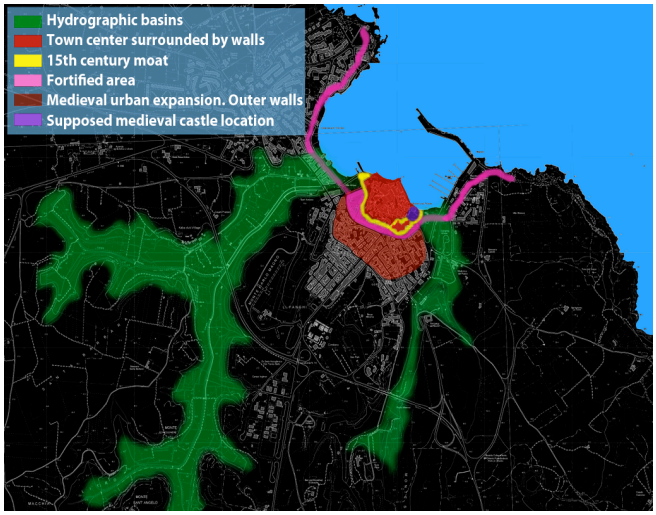



Figure 2. Definition of hydrographical, urban and defensive structures.

Without going into too much detail regarding preceding eras, for the late Middle Ages one can certainly talk of a city well defended by parallel rings of walls on the inland side and guarded by a system of towers and curtain walls on the sea, organized according to the framework usual in the poliorcetics of that time (*turres, cortinas et barbacanas* [32]).

Between the XI and XIII centuries the city did not undergo any traumatic events that influenced its form. This meant a structural continuity that substantially supported the demographic fluctuations and functional needs through constant adaptation [33]. We can suppose, then, from that period on, a certain saturation of the fabric within the inner circle.

Because of the strategic role of the port, extra-urban development occurred in such a way as to assure different levels of defence of the settlement, in order to avoid exposing large parts of the city and its resources to sacking by assailants and to impede direct attack upon the city centre.

The archaeological evidence which demonstrates the existence of an external wall built on the abutments of the ancient pre-Roman wall in medieval times supports the logic of "parallel rings" (according to a logic consistent with continuity and economic rationalism) which were built according to a byzantine model with round towers and curtain walls whose extension is still faintly visible today in the form of the development of the modern city.

Another wall or system of towers, of which there is evidence in a number of pictures, was located along the internal coast of the bay, in order to monitor for and repel eventual disembarkation by assailants; from the Swabian age, the defence of the territory was based on a rational and well developed system which involved direct or indirect communication between positions, towers and castles [34].

*2) The castle*

All this system had to have its fulcrum in the castle, the location of which during the Middle Ages is still uncertain. On the basis of descriptions of the access to the port from the sea made in the second half of the XIII century in *Lo Compasso de navegare e la Puglia* we know that the fort overlooked the sea [35].

The only descriptive reports that we have of the castrum in the first half of the 13th century refers to the necessity to leads us to imagine a fortification exposed to high tides, while at the same time, according to the *Compasso* intervene in order to repair two towers damaged by the sea (*due turres ex maris percussione continua minantur ruinam* [36]).

The present day form of the area between the port and the city, upon that we can hypothesize the medieval castle, was probably heavily modified by the excavation of the moat in the modern era. From reading the contour lines, from the signs of the quarry and the Bastion of Pelasgi, it's apparent that the original rock face was lowered by several metres, converting the original and naturally craggy slope, which was approximately 7-8 metres above sea level into the present day low lying plane which connects the 15th century moat with the sea [37], [38].

Further confirmation seems to come from the network of the urban roads (path matrix), which appears to be "oriented" towards what once must have been the 'sea gate' defence for the castle.

The constructed masses were then enclosed within a system of curtain walls interspersed with towers which opened onto the bay where the port was situated; a vital place for the economic life of the city. The castle, with its functional and symbolic value, was the fulcrum of this landscape composition.

*3) The gates*

Questions connected with the closure and defence of the urban space are tied to aspects that concern the connection of the city with the outside and with the structures of exchange.

We know from contemporary descriptions that the main gate of the city opened onto river Idro while the abovementioned seaward gate, opened towards the south [39]. Apart from the main gate, a small gate *(porticella)* connected part of the city with the surrounding countryside.

*4) The churches*

On the religious side, two ecclesiastical buildings are those best known from the archaeological investigations and studies: the cathedral [40] and the St. Peter's church [41], for which functional continuity has been established since the Early Middle Ages.

They can be considered the epicentres of well-defined sectors of the city for their importance to the cultural and iconological *meme* of Otranto.

*5) Connections and activities*

The image of the city from the land was mediated by a number of churches and monastic settlements outside the city walls, residential suburbs or facilities for warehousing goods or for the production of handcrafts and other goods. The public road that connected the city with Lecce and Brindisi to the north and with Castro and Leuca to the south skirted the external wall of the city, deviating around the city near the port. Along the basin of the river Idro, on the rocky banks, a series of caves that probably date back to the cave dwelling culture of late antiquity or the Paleochristian period

opened up [42]. Based on the principle of continual function as seen with other structures, these were probably still in use by the lower levels of the population for housing, shelter for animals, craft making workshops or as deposits for agricultural tools.

The city of the Late Middle Ages, in periods of political stability and before the Saracen raids, passed through a phase of consolidation of its economy of scale based on exchange and the port. Services and specialisation were developed supporting a well-developed social pyramid. This led to a marked diversification in the various types of housing, in large part erased by the modern walls and by more recent reconstruction.

*6) Housing*

The urban fabric within the walls presents obvious heterogeneity that is related to the peculiar stratified and paratactic condition of its formation. For the area overlooking the sea, located on a natural raised plane (acropolis), an ordinary, regular, geometric layout, based on the model of the classical Greek-Roman atrium-peristyle house which was widespread in byzantine town planning seems evident (for Otranto, upon an initial analysis of the ground floors, a base model of around 15.6x21.8 metres is revealed, corresponding to a unit measuring 50x70 byzantine feet of 31.2538 centimetres).

The reading is complicated in the lower areas where overlapping fragments of structures that resemble the Roman insulae model are positioned in order to accommodate the matrix of paths and natural terraces defined by the natural contours.

There are examples of building relating to different settlement logic. Such is the case near the cathedral and bishop's palace which is laid out on a east-west axis in correspondence with the liturgical orientation, but equally obvious is a border area which saturates the area adjacent to the linear northern front of the medieval wall, on which the Romanesque bell tower sits, then enlarged when the walls of the 15$^{th}$ century were erected.

From the XII century onwards some of these modules were replaced with terrace houses that were common in the commercial area. The basic type of structure in the medieval period, two rooms with vaulting on the ground floor and attic in wood on the upper floors [43], was based on the model of shop and residence and is found in the historical centre, in scattered agglomerations along the pathways. Other lots, originally atrium-peristyle houses were substituted in later ages (from the XVI century) to make way for the creation, in grouping or lines, of aristocratic residences.

*7) Services*

Structural reading based on the recognition of logical distribution or relative elementary modules is only one of the interpretive keys available. The city of Otranto in the Middle Ages was also characterized socially. The scene was brought to life by large or small family groups living according to a rigid subdivision and hierarchy of tasks: travellers who lived in the *xenodochia*; pilgrim beggars on their way to or from the east [44]; merchants and craftsmen; men of religion; milites and pedites.

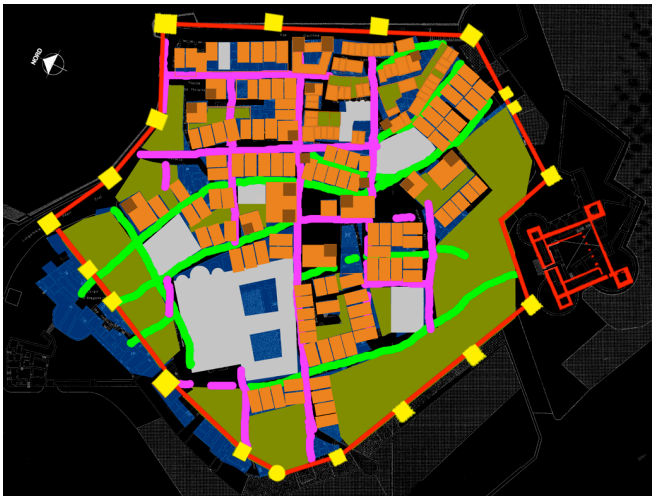


Figure 3. Schematic plan of supposed medieval Otranto.

The city itself was a machine designed for defence in case of attack, to facilitate or hinder certain categories of weapons: the windy road, narrow staircases, conferred an operational advantage on the citizens without horses who could, working from the land and their windows, fire long ranging arms) [45], etc.

All these layers overlap to define and characterise the real object of the research that opens the field to the infinity of things that can be expressed by the virtual.

In figure 3 is shown the schematic plan of supposed medieval Otranto showing the hierarchy and distribution of routes. In this figure are reported the ancient classic structure ante Middle Ages in magenta, the medieval ones according to contour lines in light green, the buildings (dwellings in orange, ecclesiastical in light grey, towers in brown), the spaces (courts in black, gardens and fields in olive green) and the fortifications (towers and gates in yellow, walls and castle in red).

The scheme is drawn on the actual ground plan of the town (in background white on black) compared to the 19th Century plan (in blue).

## VII. 3D MODELING AND GAME ENGINE

A Digital Terrain Model (DTM) that has been produced using ESRI ArcGIS, containing all historical information like sea level, rivers, etc. It has been saved in .dif format and imported in the game engine.

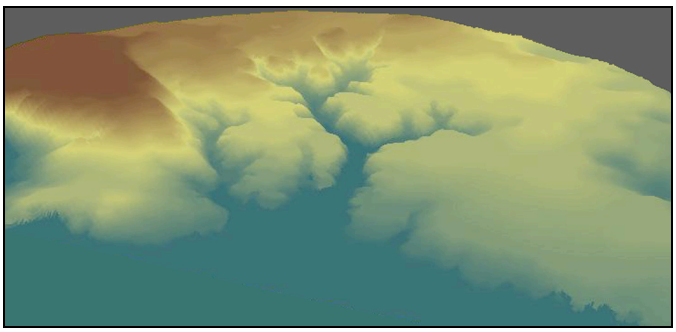


Figure 4. Digital Terrain Model of Otranto site.

Figure 5. Location of Otranto town in DTM.

In figure 4 is shown the Digital Terrain Model (with a magnification of 5) of the Otranto site and in figure 5 the location of the town in this model.

For building and street modelling, we first used AutoCAD, 3ds Max, Cinema 4D. Characters and animation are made using 3ds Max.

Once defined a list of modular elementary residential unit, according to the local medieval unit system, we composed the urban landscape in which monuments, infrastructures and situations are located.

In Figure 5 is shown the plane-volumetric reconstruction in CAD application and in Figure 6 is reported a 3D model of actual Otranto.

## VIII. BUILDING OF THE VIRTUAL ENVIRONMENT

For the building of the virtual environment we used the Torque Constructor editor of GarageGames for creating 3D architectural contents for the Torque 3D engine.

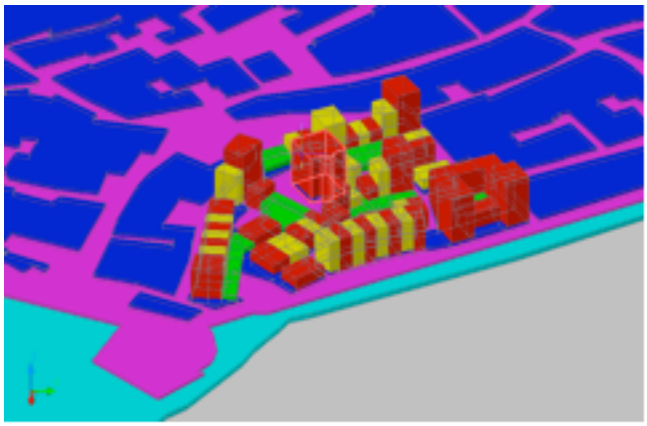


Figure 6. 3D model of buildings distribution.

For the building of specific monuments, such as Saint Peter's Church, the Cathedral and the Castle, we used first a CAM in order to obtain a more accurate definition of the architectural structures and then we imported these models into the Torque 3D engine [46].

Figure 7. Render view of actual old town of Otranto.

The choice of the Torque Constructor was prompted by technical considerations regarding the ability of software to perform a direct mapping of the files ".map", the compatibility level with the Torque Game Engine chosen to develop the game, the immediacy and the usability of internal tools. The application also includes all the converters needed to export file from '.map' to '.dif' compressed structure.

The Torque Constructor has proved to be an efficient tool for the direct implementation of 3D graphics models. In particular, it has many geometrical tools for the graphic processing of the reality context and different controls to select the top of the structure or individual brush model.

All units made in the Torque Constructor have been imported into the Torque Game Engine. The initial testing step revealed several problems of navigability of the objects. These problems were related to the incompatibility between the domains of collision associated with the objects imported into the three-dimensional environment and the avatar.

Tests carried out have helped to identify and resolve these problems by setting the values associated to the collision domains and to the proportions between objects and avatars. At present all units are properly imported and successfully navigated.

In Figure 4 a set of residential units are shown.

In the context of the computer graphics for cultural heritage, a stable algorithm has been implemented to import CAD objects into the Torque Game Engine platform and to ensure navigation into each graphic structure. This technique together with an efficient system for exporting textures and paintings will be used to realize graphic complex environments for the 2D/3D reconstruction in cultural heritage.

The first monument to be modelled has been St. Peter's Church, due both to its characteristic of modularity that is useful for testing the software and its historical relevance as unique byzantine building located in a medieval context.

After drawing and importing the church with textures and lights we experienced problems with the non-convex objects produced by common modelling software that drove us to use only Torque dedicated applications like Torque Constructor.

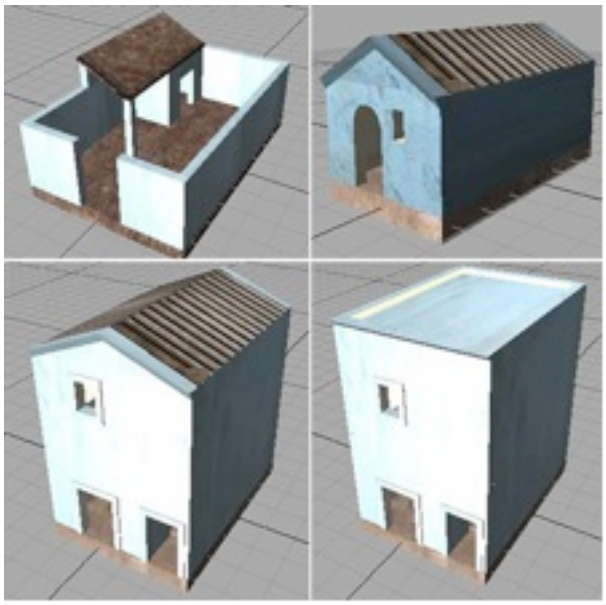


Figure 8. A set of residential units

In Figure 11 is shown the reconstruction of St. Peter's Church and its surroundings; in particular, in 11(a) is reported the scheme of the reconstructed church with (in black) a chapel that existed in the Middle Ages and was afterwards destroyed.

In Figure 12 is shown the reconstruction of Otranto Cathedral; in particular, in 12(a) is reported the mosaic of the internal floor of the church.

## IX. PLAYERS AND ARTIFICIAL INTELLIGENCE

Inside MediaEvo Project has been implemented a module to manage the interactions with Artificial Intelligence [47]. The artificial intelligence (AI) is necessary to establish relations among characters in the virtual game and to exchange multimedia information and by prompting commands real time. The ability to interact with AI characters is the principal key for retrieving knowledge and experiences from a virtual reality environment.

In the MediaEvo Project, the component of Artificial Intelligence is based on a graphical interface, with the following specifications: the interface should allow the starting of the interaction by pushing a default button on the keyboard; the interface should provide a choice of applications to be given as instructions to the virtual character; the interface should display all workable interactions with a virtual character. For this purpose, a reconfigurable database of instructions has been generated.

The configurable database has direct access to the AI Interactive module. The AI Interactive Module has been realized according to the guidelines of the scripts implemented in Torque Game Engine [47].

In figure 9 is reported the algorithm to manage the Artificial Intelligence.

The AI Interactive algorithm can be divided into two main modules: AIT Server Management Code and AIT GUI Management Code. When the player selects an item of the AIT Queries database, the GUI interface establishes a communication between the player and a virtual AI character. The selected item contains the instruction that could be imparted to the AI character. The instruction is managed from the AIT GUI Management Code module that encapsulates the information into a single system call.

Finally, the system call is routed to the AIT Server Management Code module and then it is interpreted to identify the corresponding action, into the AIT Actions database.

The game has been designed for enabling multi-playing in order to provide a real-time interaction with other game sessions localized in the reconstructed virtual environment.

Some multimedia elements are available in the MediaEvo platform for the context of edutainment in cultural heritage. The main ones are: 1. the availability of audio clips and sounds in the game; 2. the use of triggers to start up audio or video events when a player reaches some checkpoints or thresholds.

In Figure 10 is shown the visualization of a video that pops up when a player gets close enough to the entrance of St. Peter's church. In the same figure 10 is also shown a virtual radar, a mean to let players know their position in the town and the one of other players.

## X. THE ALGORITHM FOR THE AUDIO AND VOICEIP

Inside the MediaEvo Project has been implemented a module to integrate voice and text interactions between the players and other characters located into the Torque virtual environment.

The trigger between the vocal process and the Torque virtual environment is realized through a system call implemented and built into the kernel of the Torque Game Engine.

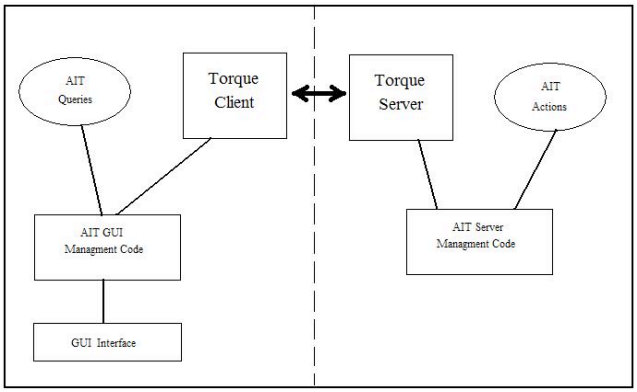



Figure 9. Artificial Intelligence.

Figure 10. The opening of a multimedia clip video.

The insertion of the vocal connections can increase the interactions with characters and/or players in the virtual game and allow performing some input and output that normally are performed throughout keyboard, mouse, screen, and other input/output devices.

The possibility to establish a vocal connection with other players is one of the best ways for retrieving knowledge and experiences during the virtual game. The vocal interaction contains some algorithms to realize text-to-speech systems. It is also possible to ask some actions directly to the players through the audio channel.

In addition, in the audio module has been implemented a system to realize a VoiceIP connection between all players. The audio module is based on a simple graphical interface provided with a bar that indicates the microphone audio level during the vocal conversation and a flag that specifies if the audio module is working or not.

The audio module has been integrated with some scripts inside the Torque Game Engine. All the audio conversations are transmitted through a protocol compatible with the Internet platform.

The Speaky toolkit [48] has been provided by MediaVoice Company, partner of the MediaEvo Project, and is based on two modules: the Voice Platform and the Control Center Modules.

The main task of the Voice Platform is to handle the voice interaction between the user and the applications. The Control Center helps the users to configure Speaky parameters.

The Speaky platform supports Loquendo engine for Automatic Speech Recognition (ASR) and Text To Speech (TTS) in Italian language. The MediaVoice Company is working for a multi-language version of the product.

The vocal interactions are realized by using a specific remote command that can communicate with the Speaky Toolkit for imparting voice interactions.

## XI. CONCLUSION AND FUTURE WORK

The MediaEvo Project has led the researchers to consider some of the issues presented by the multidisciplinary nature of the project and the close correlation between technical and humanistic fields. In particular, conditions were created that implied history researchers to test model built using information coming from their work.

In larger terms, the reconstruction of the medieval city of Otranto in the MediaEvo Project, determined the conditions for testing the overall functionality and systemic coherence through the real time production of environments, objects, situations, and virtual landscapes, thought up in order to represent the totality of knowledge of that times.

The representation, intended for communicating and educating, was designed to open itself up to the interactive and multisensory dimension in such a way as to become simultaneously subject, object and context of the experience. Communication and representation are not limited to the pathway of a one-way guided narrative, but open up possibilities for more enjoyable elements of interaction and a multisensory mediation, in which can merge objects, subjects and the experiential context.

Measured against the notable potential of a virtual scenario, a series of properties have been defined sufficiently to give the game platform an effective educational value [49].

By incorporating historical, technical and educational considerations the final product presents itself as a “complete-open-interactive” environment, with a good historical-philological validation, while allowing for continuous updating and testing. Since the Middle Ages are only partially explored, by these means an ideal extent of the knowledge of a context can be represented and experienced in its material totality.

Already tested for other urban realities, by opening itself to exponential complexity, the Time Machine is definitively becoming a formidable tool for the acquisition of knowledge, the enhancement and safeguarding of cultural heritage.

The MediaEvo Project evaluates the premises upon which the future development of an historical cyberspace is capable of contextualising past experience, in order to explore a range of parallel realities based on description and philological reconstructions.

In the MediaEvo Project has also been tested the possibility to enjoy the virtual environments using the Apple IPhone mobile. The IPhone version of MediaEvo Project through the iTGE platform for IPhone Torque is still in progress.

## ACKNOWLEDGMENT

The authors wish to thank Pierpaolo Limone of Foggia University for the advice in the pedagogical development of the game and Massimo Limoncelli for the building of the 3D model of the Otranto Cathedral.

## REFERENCES

[1] T. R. Gruber, "A translation approach to portable ontology specifications", in Knowledge Acquisition, vol. 5, issue 2. june1993, pp.199-220.

[2] R. A.Rosenstone,. Revisioning History. Film and the Construction of a new Past, Princeton, USA: University Press, 1995.

[3] J. Baudrillard, "Pataphysique de l'An 2000", in: L'Illusion de la fin ou La grève des événements. Paris, France: Galilée, coll. L'Espace critique, 1992, pp. 11-22.

[4] D. De Kerckhove, Brainframes: Technology, Mind and Business, Baarn, The Netherlands: Bosch & Keuning,. 1991

[5] F. Morganti, G. Riva, Conoscenza comunicazione e tecnologia. Aspetti cognitivi della realtà virtuale, Milano, Italy: LED Edizioni Universitarie di Lettere Economia Diritto, 2006.

[6] C. Borgatti, L. Calori, T. Diamanti, M. Felicori, A. Guidazzoli, M.C. Liguori, M.A. Mauri, S. Pescarin, and L. Valentini, "Databases and Virtual Environments: a Good Match for Communicating Complex Cultural Sites", Proceedings of ACM SIGGRAPH 2004, Los Angeles, USA, 2004.

[7] S. Pescarin, Reconstructing Ancient Landscape. Budapest, Hungary: Archaeolingua, 2009, pp. 21-23.

[8] R. Dawkins, The selfish gene. Oxford, UK: University Press, 1976.

[9] M. Forte, "Mindscape: ecological thinking, cyber-anthropology, and virtual archaeological landscapes" in M. Forte, & P.R.Williams (Eds.), The reconstruction of Archaeological Landscapes through Digital Technologies, Proceedings of the 1st Italy-United States Workshop 2001. Boston, USA, 2003, pp. 95-108.

[10] M. Song, T. Elias, I. Martinovic, W. Mueller-Wittig and T. K.Y. Chan, "Digital heritage application as an edutainment tool", Proceedings of the 2004 ACM SIGGRAPH International Conference on Virtual Reality Continuum and its Applications in Industry, Singapore, 2004, pp. 163–167.

[11] P. Kiefer, S. Matyas and C. Schlieder, "Learning about Cultural Heritage by playing geogames". Lecture Notes in Computer Science, Springer Berlin/Heidelberg, vol. 4161/2006, Book Entertainment Computing, 2006.

[12] G. Cutrí , G. Naccarato and E. Pantano, "Mobile Cultural Heritage: the case study of Locri", Lecture Notes in Computer Science, Springer Berlin/Heidelberg, 2008, pp. 410-420.

[13] K. Luyten, J. Schroyen, K. Robert, K. Gabriëls, D. Teunkens, K. Coninx, E. Flerackers and E. Manshoven, "Collaborative gaming in the Gallo-Roman museum to increase attractiveness of learning cultural heritage for youngsters", Second Intern. Conference on Fun and Games 2008, Eindhoven, The Netherlands, October 20-21, 2008.

[14] Istituto per i Beni Archeologici e Monumentali, http://www.ibam.cnr.it/

[15] C.D. Fonseca, D. Roubis, and F. Sogliani (Eds.), Jure Vetere. Ricerche archeologiche nella prima fondazione monastica di Gioacchino da Fiore (indagini 2001-2005). Cosenza, Italy, Rubettino, 2007, pp. 87-132.

[16] Project Itinera, http://www.itinera.puglia.it/

[17] M. Forte, S. Pescarin, E. Pietroni, "The Appia Antica Project", Proceedings of the 2nd Italy-United States Workshop, Berkeley, USA, May 2005.

[18] Virtual Rome Project, http://3d.cineca.it/storage/demo_vrome/htdocs/

[19] Muvi, http://muvi.cineca.it/

[20] F. Bocchi, "The city in four dimensions: the Nu.M.E. Project" Journal of Digital Information Management. Vol. 2, issue 4, 2004, pp. 161-163.

[21] City Cluster, http://www.fabricat.com/CITYCL_WEB2003/CITYCLUSTER.html

[22] Quest Atlantis Project, http://atlantis.crlt.indiana.edu/

[23] Integrated Technologies of Robotics and Virtual Environment in Archaeology Project, Virtual Heritage Lab, http://www.vhlab.itabc.cnr.it/FIRB/Release/Home.html

[24] T. Salmon, Lo stato presente di tutti i Paesi e Popoli del mondo naturale, politico e morale, vol. 23, Napoli, 1763.

[25] H. Houben, Otranto nel Medioevo: tra Bisanzio e l'Occidente, Galatina (Lecce), Italy, Congedo, 2007.

[26] H. Houben, La conquista turca di Otranto (1480) tra storia e mito, Galatina (Lecce), Italy:, Congedo, 2008.

[27] D. Michaelides, D. Wilkinson, Excavations at Otranto, The excavation, Vol. I, Galatina (Lecce), Italy, Congedo, 1992.

[28] .D'Andria, D. Whitehouse, Excavations at Otranto. The Finds, Vol. II, Galatina (Lecce), Italy:, Congedo, 1992.

[29] G. Caniggia, G. L. Maffei, Lettura dell'edilizia di base, Venezia: Marsilio, 1979.

[30] S. Previtero, "Osservazioni sulla metrologia antica e medievale nel Salento", in S. D'Avino, M. Salvatori, Eds., Metrologia e tecniche costruttive, Contributi, vol. 5, 1998, p. 97.

[31] Y. Renouard, Le città italiane dal X al XIV secolo, Torino, Italy: BUR, vol. I, 1975, p. 16.

[32] D. Vendola, Ed., Documenti tratti dai Registri Vaticani. Da Innocenzo III a Nicola IV, vol. I, Trani (Bari), Italy: Vecchi, 1940, n. 336, p. 263.

[33] Anonymus Barensis, Chronicon, edited by L. A. Muratori, in Rerum Italicarum Scriptores, vol. 5, Milano, Italy, 1724, p. 152.

[34] R. Licinio, Castelli medievali. Puglia e Basilicata: dai normanni a Federico II e Carlo d'Angiò, Bari, Italy: Dedalo, 1994, pp. 117-130.

[35] O. Baldacci, "Lo Compasso de navegare e la Puglia", in Annali della Facoltà di Magistero dell'Università di Bari, Bari, Italy, vol. 2, 1960, pp. 200-201.

[36] E. Sthamer, Ed., Dokumente zur Geschichte der Kastellbauten Kaiser Friedrichs II und Karls I von Anjou, 2: Apulien und Basilicata, Tübingen 1997, n. 1014, p. 155.

[37] H. Goosse, O. Arzel, J. Luterbacher, M. E. Mann, H. Renssen, N. Riedwyl, A. Timmermann, E. Xoplaki, and H. Wanner, "The origin of the European Medieval Warm Period", in Climate of the Past Discussions 2, vol. 3, 2006, pp. 285-314.

[38] F. Ortolani, S. Pagliuca, "Evidenze geologiche di variazioni climatico-ambientali storiche nell'Area Mediterranea", in Quaderni della Società Geologica Italiana, vol. 1, march 2007, pp. 13-17.

[39] M. Amari, C. Schiaparelli, Eds., L'Italia descritta nel "Libro del Re Ruggero" compilato da Edrisi, Roma, Italy, 1883, p. 76.

[40] S. Mola, R. Cassano, M. Pasculli Ferrara, "La cattedrale di Otranto", in C.D. Fonseca, Ed., Cattedrali di Puglia. Una storia lunga duemila anni, Bari, italy: Adda, 2001, pp. 237-243.

[41] L. Safran, S. Pietro at Otranto: Byzantine Art in South Italy, Roma, Italy: Rari Nantes, 1992.

[42] G. Uggeri, "Otranto paleocristiana", in Itinerari (contributions on the history of art in honour of Maria Luisa Ferrari), vol. I, Firenze, Italy: S.P.E.S, 1979, pp. 37-46.

[43] G. Strappa, M. Ieva, M.A. Dimatteo, La città come organismo. Lettura di Trani alle diverse scale, Bari, Italy: Adda, 2003, pp. 69-71.

[44] "Vita Beati Nicolai Peregrini (B.H.L. 6223) ", in O. Limone, Santi monaci e santi eremiti. Alla ricerca di un modello di perfezione nella letteratura agiografica dell'Apulia normanna, Galatina (Lecce), Italy: Congedo, 1988, p. 143.

[45] E. Guidoni, L'arte di progettare le città, Italia e Mediterraneo dal Medioevo al Settecento, Roma, Italy: Kappa, 1992.

[46] K. C. Finney, Advanced 3D game programming all in one, Boston, USA: Thomson Course Technology, 2005.

[47] K. C. Finney, 3D game programming all in one, Boston, USA: Thomson Course Technology, 2004.

[48] Speaky toolkit, MediaVoice srl, http://www.mediavoice.it

[49] Oliva L., De Paolis L.T., Aloisio G., "Otranto nel medioevo. Ricerca e ricostruzione urbana per l'edutainment", in REM, Rivista ufficiale della SIREM, Società Italiana di Ricerca sull'Educazione Mediale, vol. 1, issue.2, december 2009, Trento, Italy: Erikson, pp. 199-212.

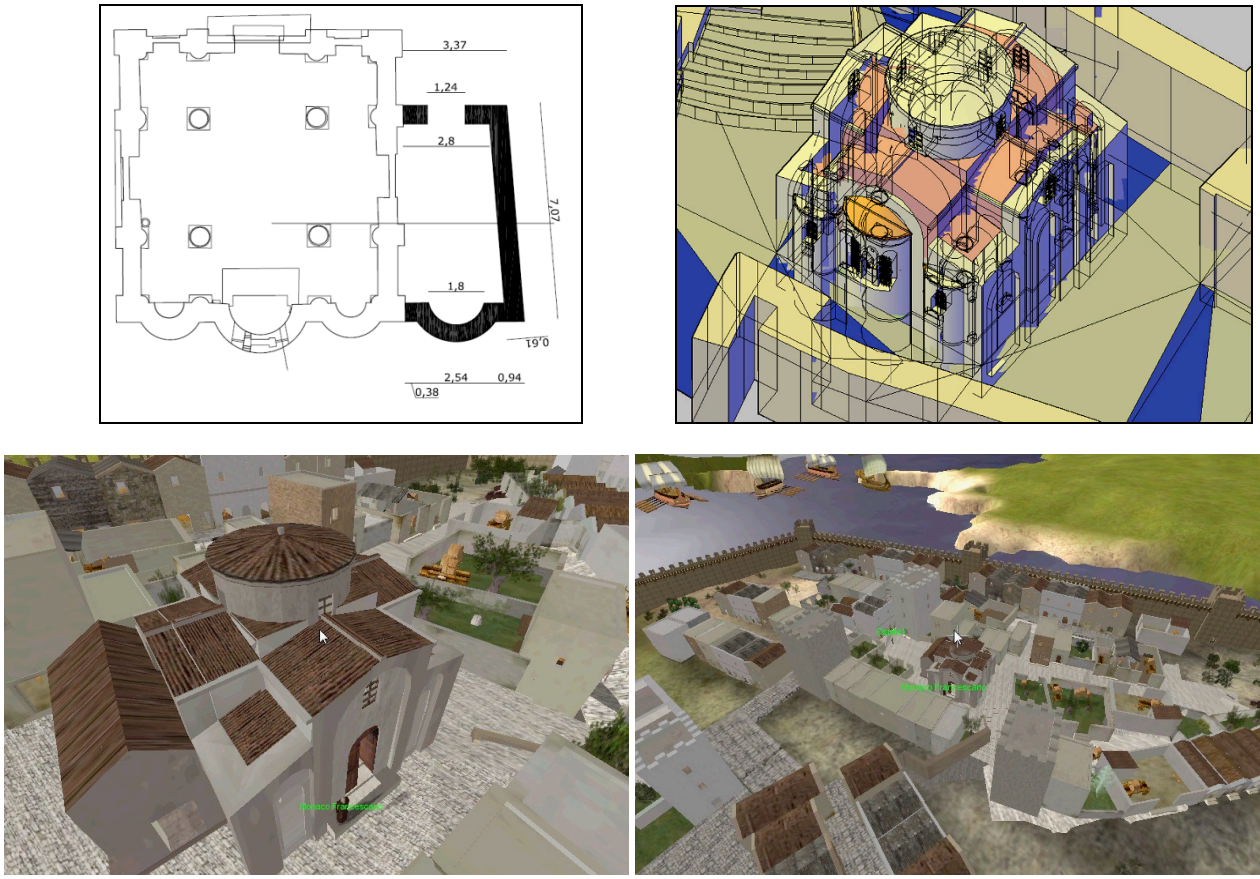



Figure 11. The reconstruction of St. Peter's Church: (a) scheme of the reconstructed church with the later removed chapel in black; (b) virtual model made using a CAD software; (c) the virtual reconstructed church; (d) the church in its surroundings.

Figure 12. The reconstruction of Otranto Cathedral. 12(a) facade, exterior view; 12(b) the famous medieval mosaic on the internal floor.

# Mobile Robot Localisation and Terrain-Aware Path Guidance for Teleoperation in Virtual and Real Space

Ray Jarvis
Intelligent Robotics Research Centre
Monash University
Victoria, Australia
ray.jarvis@monash.edu

***Abstract*–This paper concerns the development of a force feedback enhanced teleoperation system for outdoor robotic vehicles navigating in rough terrain where true-colour 3D virtual world models of the working environment, created from laser and colour image scans collected offline, can be explored by walk-throughs both before and during the robot navigation mission itself. In other words, the physical mission intended can be partially rehearsed in cyberspace[1]. Further, during a mission, the location and orientation (localisation) of the vehicle are continually determined and global collision-free paths to selected goal locations made available as advice to the operator, who can follow or ignore such advice at will. Live (real-time) 3D laser range data also provides an up-to-date scan of the volume immediately surrounding the vehicle as it moves so that dynamic obstacles can be avoided. Local terrain-roughness is taken into account in the provision of local collision-free paths, the sub-goals of which are operator determined. This live range data is matched with the pre-scanned range data to calculate the accurate robot vehicle localisation (position and orientation) which is provided continuously during the navigation mission. A force feedback 3D joystick reflects terrain roughness as a vibration in one axis and the other two axes are used to provide a 2D force to attract the operator towards following the local optimal collision-free path, but this attraction can be easily overridden by the operator. The instrumentation and methodologies used for localisation, path planning, force feedback teleoperation and 3D exploration are presented, together with some preliminary experimental results for large outdoor, natural environments.**

***Keywords-Human/Machine Interaction, Teleoperation, Localisation, Cyberspace, Robot Navigation, Rough Terrain, Force Feedback.***

## I. INTRODUCTION

In the realm of mobile robot navigation, the research community has long held fully autonomous operation as the ultimate goal. Yet, in many practical situations, this is not currently possible and, in some, not really justifiable or even sought after. Two examples where fully autonomous robot navigation is ether not sought for or infeasible are provided as follows:

A severely disabled patient may be reliant on wheelchair navigation for his/her mobility needs [2]. Whilst providing sensor-based obstacle avoidance and safe-path guidance may contribute to the user's capacity to better engage the world of mobility, fully automating the process would impinge upon that person's freedom and also cause some reduction of capacities supporting independence still held to be of value in a quality of life sense. The second example could be in a bush fire fighting situation [3], where an operator is available to provide human judgement and mission sub-goals but should not be in risk of physical injury or death. Sensor informed feedback based teleoperation would suit that situation well. Again, some navigation support would be welcomed but full automation not really required (nor currently feasible).

This paper concerns remote teleportation of robotic vehicles, possibly in fire- fighting or search and rescue operations in outdoor rough terrain situations, with sensory feedback and path guidance support. The manner in which the human agency interacts with the system and interprets newly developing situations is considered critical to the quality of the navigation in the context of higher level mission goals.

Robot navigation systems have three essential components and several more peripheral ones. Firstly, the location and orientation (pose) of the robot vehicle needs to be known in the context of its current working environment. This is known as

'localisation'. This can be geometrical or topological in nature and may depend on the recognition of man-made or natural landmarks. Various instruments such as global positioning systems (GPS), flux-gate compasses, wheel odometry, video cameras, laser range finders and inertial systems can be employed for this. The second requirement is the availability of a map of the working terrain or the means of acquiring one whilst navigating. In recent times, considerable research effort has been expended on simultaneously localising the robot and developing an environmental map (SLAM-Simultaneous Localisation and Mapping). There are a number of difficulties using SLAM in the context of the application considered in this paper. These will be touched upon later. The third requirement is collision-free, low risk and somewhat optimal path planning. Ideally the terrain properties, including roughness as well as obstacle structures, should be taken into account by the path planning strategy.

In the SLAM approach [4,5], the environmental map takes quite some operational time to construct and optimal path planning cannot take place before the completion of the map, although piece-wise optimal strategies can be implemented within the context of partially known environmental spaces. There are also some problems with reliably recognising closures (places revisited) to distribute accumulated errors optimally.

In this paper, an alternative approach has been adopted -that of acquiring, off- line, a detailed and accurate environmental map before, perhaps one of many, robot navigation missions are executed. It is admitted that this may not be always practicable but, for many situations, the collecting of the map data can be treated like any other preparation step in anticipation of a crisis scenario which may eventuate later. Clearly, for urban environments which could be subject to natural disasters like fires, floods and earthquakes, this precaution is very reasonable. In bushland settings near homesteads this could also be seen as feasible. Even entire farms with forest stands subject to fire risk could be pre-scanned in this way. Scanning instruments with quite large operational volumes are currently available. These are somewhat expensive, but one could imagine a bureau service providing the scan data for an affordable fee and even insurance companies reducing premiums for clients who have obtained this data. Besides, this technology will become less expensive with time.

The remainder of this paper is structured as follows. The next section describes, briefly, a number of outdoor vehicles instrumented for teleoperation as part of a research effort supporting bushfire fighting. Any one of them could be operated using the navigation system which is the subject of this paper. Next, the instrumentation, both for off-line mapping of the environment and the on-board real-time laser range scanning, which are crucial for this work, is described. Then, a section on localisation and path planning using the results of scanning follows. The whole navigation system with force feedback for assisted teleoperation is then introduced. Discussion and future work follows prior to the conclusions section.

## II ROBOTIC VEHICLES

Figure 1 shows a number of standard (commercially available) vehicles which have been instrumented for teleoperation as part of a research project to support bush fire fighting, where the local Country Fire Authority (Victoria, Australia) was the industry partner. The variety of vehicles represents a number of different, yet related, activities supporting bush fire fighting. A four wheel drive farm 'bike' fitted with tracks [Figure 1(a)] is capable of climbing over fallen tree trunks up to 40cm thick and has been targeted mainly for forward scout forays to assess the severity and access possibilities along fire-break tracks prior to fire fighting itself [6]. It can also be used for very rough terrain search and rescue for firemen and property owners who may have became asphyxiated or have suffered smoke blindness. The heavy tracks are extremely difficult to steer and a powerful chain linked hydraulic ram system has been employed for changing the steering direction of the front tracks. Steering, braking and acceleration can all be operated by remote control via standard

'hobby' style servo actuators and a radio control transmitter/receiver pair or, alternatively, by computer Ethernet links to serial line servers which can operate the servo actuators. An excavator [Figure 1(b)] and a front loader [Figure 1(c)] are also teleoperable and are targeted for fire-break track clearing and smoothing for fire tanker access [7]. In both cases, in addition to mobility controls (steering, brake, accelerator), the buckets can also be teleoperated. Figure 1(d) shows a 40 foot boom truck which can used both for search and rescue, with high vantage point views, and the capability of lifting a human up from behind a wall of fire and for directing a steam of water from the boom bucket [8]. Finally [Figure 1(e)] there is a fire tanker [2] which can have 3000 litres of water and spray it at selected directions using a pan/tilt device aiming the water flow [Figure 1(f)].

Whilst all the above vehicles can be fitted with video and infra-red video cameras and laser range finders to assist teleoperation, the particular laser range finder instruments described next are the specific devices which support the main emphasis of this paper.

Fig. 1(a). Four-Wheel Drive Farm


Fig. 1(b) Excavator


Fig. 1(c) Front Loader

Fig. 1(d) 40 foot Boom Truck

Figure 1(e) Fire Tanker

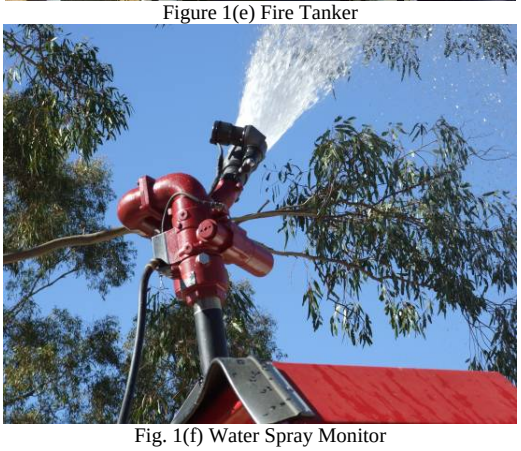


Fig. 1(f) Water Spray Monitor

FIG. 1. ROBOTIC VEHICLES

## III. CRITICAL LASER RANGE FINDER SCANNERS

Two distinct laser range finder instruments are crucial in their support of this research. The first collects pre-mission environmental data (range and colour) to build an accurate 3D cyberspace of the working environment and the second collects real-time range data during the navigation mission itself.

A Riegl LMS-Z420i [see Figure 2(a)] is an accurate time-of-flight laser range finder which can be fitted with a high resolution digital camera whose image data can be registered with the range data. This instrument can range up to 800 metres with an accuracy of 1cm, collecting range values at up to a 11,000 samples per second rate. A typical medium density scan from a fixed position takes between 15 and 60 minutes, depending upon the settings used. Since not all aspects of a 3D scene are viewable from only one location, several fairly open locations are chosen for individual scans and these are later fused together under human supervision with computational support. These separate scans should overlap to allow accurate registration during integration. A two metre diameter 'dead zone' exists around the instrument since, up to this distance, the return timing is too short for the instrument to record correctly. A typical view of a scanned space is shown in Figure 2(b).

The second laser range finder is a Velodyne HDL-64E S2 [see Figure 3(a)] which spins at a rate of 5-15 Hz to collect range data up to 120 metres away (dependant upon the target surface albedo) at a date rate of up to 1.8 million samples/second at an accuracy of 2cm.

The Velodyne contains 64 independent laser sources and sweeps 64 live scans around the axis of rotation, collecting data from +2° to -24.8° in elevation. When mounted high on a vehicle it allows the volume that vehicle can move through to be analysed for obstacles and also permits the terrain undulations and holes to be analysed. A typical scan is shown in Figure 3(b).

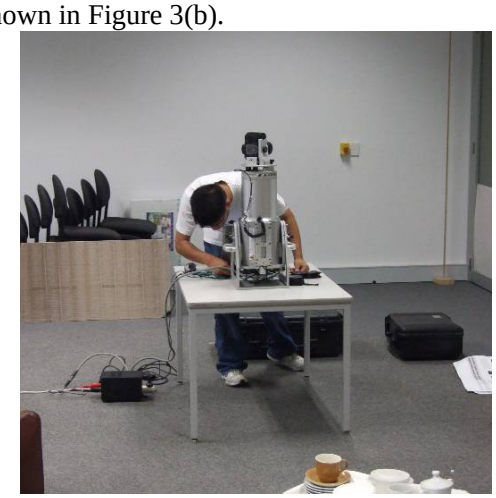


Fig. 2(a) Riegl Scanner

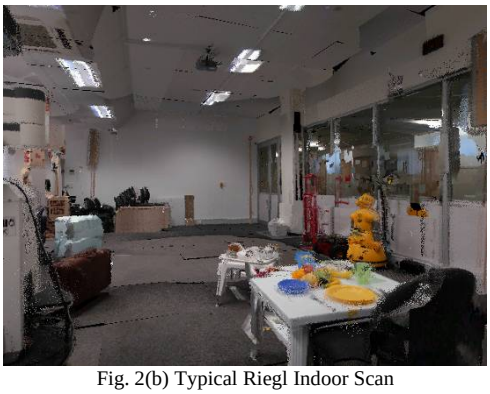


Fig. 2(b) Typical Riegl Indoor Scan

Fig. 3(a) Velodyne Range Scanner

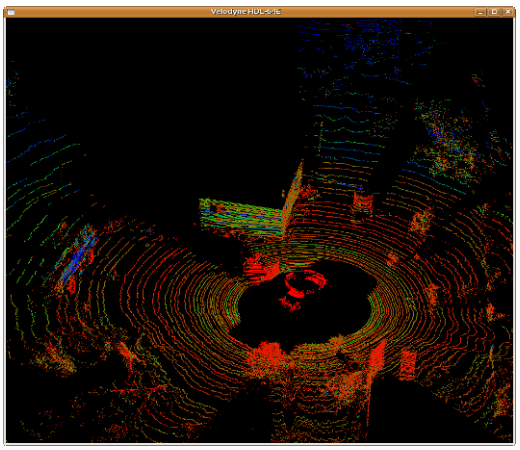


Fig. 3(b) Velodyne Outdoor Scan Example

Both instruments can be powered by standard 12Volt batteries and are connected to the controlling computer via Ethernet, but with the digital camera requiring a USB port.

The range and colour imagery collected by the Riegl scanner and attached digital camera at various locations and subsequently combined, can be explored as a virtual world, moving, through it at ground level or from a 'fly over' elevated view. This exploration can be used for pre-mission familiarisation and for noting specific aspects such as the location of dwellings or water sources, fences, gates etc. which may assist in the mission itself. It can also be used to make judgements on tolerances for obstacle avoidance which should be used during the mission and where grass and bush may be navigable despite perhaps being regarded as obstacle space because of its height.

## IV. LOCALISATION AND PATH PLANNING

The knowledge of the pose (position and orientation) of the robotic vehicle is an important requirement for efficient path planning and following, even if it were not strictly necessary for teleoperating a vehicle using on-board sensors alone (eg. cameras and range finders).

A data-base of range 'signatures' is first extracted from height thresholded (between 0.5 and 1.0 metres) range data from the Riegl scanner at intervals over a 0.1 metre grid over the working environment, associating the range to obstacles of 180 radial rays at 2° intervals around the 360° sweep, with each ray length larger than 50 metres marked as 0 (keeping only values clearly within the range scope of the Velodyne). This data-base is constructed off-line so its computational time cost is not crucial. Some local averaging is done to smooth the data to enable better spatial matching tolerances. In real-time, whilst the robotic vehicle is navigating, a similarly constructed 'signature' from height thresholded Velodyne range data and matched through searching the 'signatures' in the data base is used to determine the pose of the vehicle. A rough match is followed by a more refined one to improve the efficiency of the method. The robot vehicle can be localised, typically, within ~15cm of its actual

location and ~1 degree of its actual orientation at the rate of 0.35 seconds per fix using a fast Intel i7 2.67 GHz processor with 6 Gb of RAM. Continuity constraints are used to limit the search requirements once the vehicle is initially localised, a complete initial search taking a number of seconds.

The simple matching formula used is as follows: Given two 'signatures', one extracted from the current Velodyne range scan and one selected from the Riegl pre-scanned data base, S1 and S2, respectively, each with 180 range components.

S=Sum[exp(-Abs(S[i]-S2[i])**2/(2*Sig)] over i=1 to 180 where Sig is a experimentally selected standard deviation and Abs the absolute value operator. This produces a Gaussian weighted measure which downplays badly matching range rays.

Then X= w+b*S where w and b are experimentally determined parameters. The final score is calculated by Score= 1/(1+exp (-X)) which is between 0.0 and 1.0. The larger the score, the better the match. Further details can be found in [9].

Clearly, more sophisticated matching techniques can be developed but this first approach was found adequate for our purpose, since the terrain we used in our initial experiments was reasonably planar. The pose data (position plus orientation) is exported continuously to a text file for the path planner to access when necessary, the most recent information overwriting the previous pose data. Figure 4 shows a coherent sequence of localisation traces (in real space with a physical vehicle) with the current location for each 'screen shot' showing the Velodyne range rays which were matched against the pre-scanned Riegl data to determine the location/orientation of the vehicle. The view of the cyberspace model obtainable from that point is also shown from ground level. One can identify correspondences between objects in that view and some structures in the plan map showing the localisation point. The vehicle is approaching a shack with a fire tanker (red) looming larger. The smoothness and continuity of the traces is clearly impressive and indicates a very high confidence in the reliability and accuracy of the methodology used.

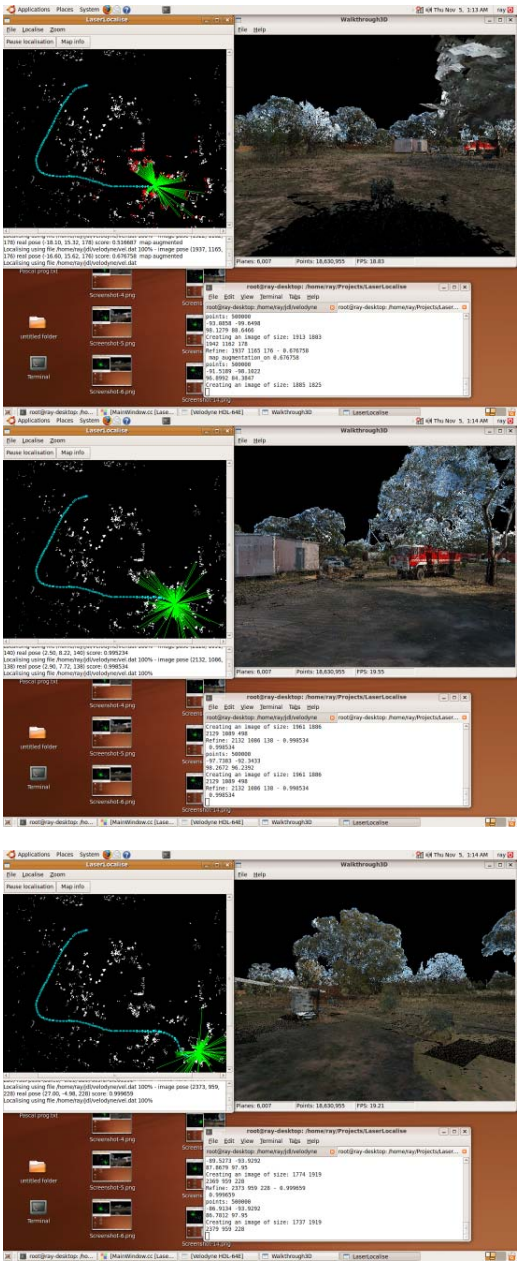


Fig. 4 Sequence of Localisation Traces and Virtual Reality Viewpoints for an Actual Physical Experiment.

A number of path planning methodologies have been published [10,11,12]. Many treat the search space as a Euclidean geometry domain made up of points and lines with polygonally enclosed obstacle spaces. Details can be found elsewhere [10]. An alternative approach is grid based, where the search space is made up of tessellated (generally rectangularly) cells which are either occupied by obstacle or not (free). A path in such a space is a sequence connected free cells form a start cell to a goal cell. The computational burden of such methodologies is highly related to the resolution chosen for the environment space representation. A big advantage of the grid cell based approach is that, in addition to occupancy or not of obstacles, other cost structures can be represented in the cellular structure so that properties such as visibility or terrain roughness etc. can be accommodated in the path optimality calculations. One can even include tolerance costs in relation to the proximity of obstacles so as to allow the robot to stray off its path to some extent without collision.

A Distance Transform (DT) path planning strategy was used in this study as it has a number of advantages which suited the needs of the project [13] despite there being more recent and complex alternatives. It is simple to compute, can accommodate costs over the cell structure, including collision risk tolerance and probabilistic structures and can easily be extended into time/space for both deterministic and probabilistically estimated cost structures projected into the future. It can include multiple goals and provides an optimal path from any cell in free space to the least cost acquirable goal simply by following a steepest descent trajectory in the DT space. This last properly is particularly useful, since, if the robotic vehicle is driven off the currently mapped out path, a new optimal path from its new position is instantly available using a new steepest decent trajectory in the already calculated DT space. The details of the DT method can be found elsewhere [13] but an outline is provided here for completeness and for better being able to explain the path-guided teleoperation approach which is described later.

First consider the simple case of an initially rectangularly tessellated N x N cell space with free cells marked '0' and obstacle cells marked '1' with only one goal.

1. Leave the goal as '0', putting a large number in all other free-cells (say > $N^2$) and mark the obstacle cells with computer infinity (say $2^{32}-1$).
2. In raster order (left to right, top to bottom, one step at a time), skipping over obstacle cells, replace the free cell value with the least value (cost) of recently visited neighbours (3 x 3 region) plus 1 (assuming that costs from entering the cell from any of its neighbours to be identical). In fact only 4 comparisons are needed (three cells in the previous line and the one to the left) but all can be used without error. The goal cell should not be altered as it is zero cost from itself.
3. In reverse raster order (right to left, bottom to top, one step at a time) repeat the operation described in 2. Now only the cells in the line below and to the right need to be looked at.
4. Repeat 2 & 3, above, alternatively until no further changes occur.
5. The resulting map is the Distance Transform and a steepest descent trajectory from any free-cell will lead to the goal with the least number of steps

Some border conditions need to be set so the rasters, are usually carried out over a (N-1) x (N-1) grid.

A simple example of a DT result is shown in Figure 5(a). If the cost of a diagonal move is preferred to be $\sqrt{2}$ compared to a up/down or left/right cost of 1, the approximation of a weight of 3 for diagonal moves and 2 for the others can be used [see Figure 5(b)]. In fact, 4:3 is even better and 17:12 almost perfect. In this case the candidate cell value is replaced by the least value of the sum of its neighbour's cost plus the cost of entering the cell from that neighbour. If costs are to reflect distances as well as roughness, tolerance or probabilities, the

same process can be used, as long as all costs are non-negative. No local entrapment occurs using this strategy and the paths formed by steepest descent trajectories are truly global at all times. Only unreachable cells (enclosed by obstacle cells) are indeterminate.

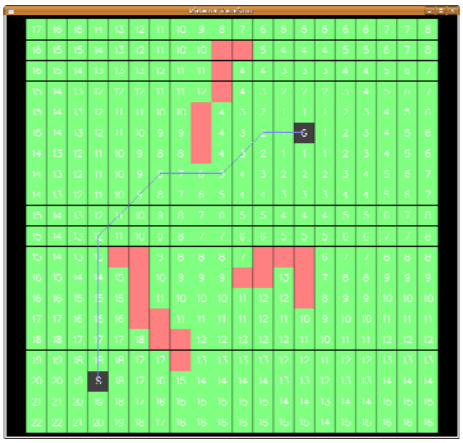


**Fig. 5(a) Simple DT Result**

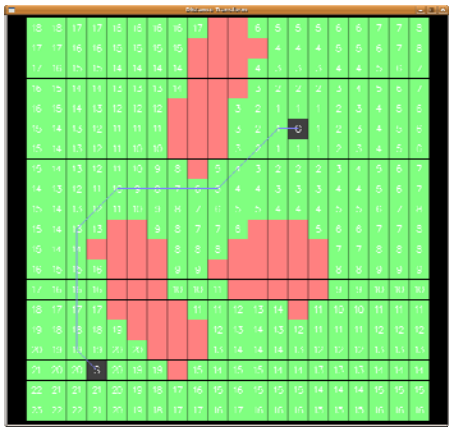


**Fig. 5(c) Grown Obstacle Field**

A particularly elegant way of 'growing' obstacles to increase collision-free tolerance and/or to allow for the physical dimensions of the robotic vehicle, is to initially treat all obstacle cells as pseudo goals (set to 0) and carry out the DT computation which leaves all free-cells with values equal to their distance from their nearest obstacle cell. Returning all values larger than a set threshold (say equivalent approximately to the radius enclosing

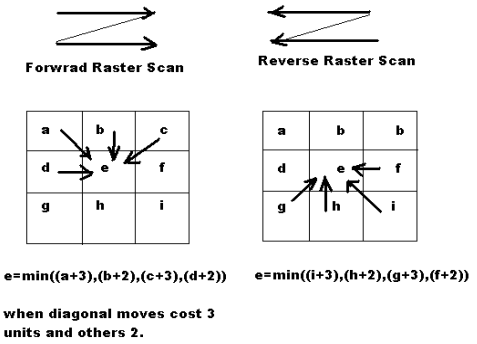



**Fig. 5(b). Raster Ordering and Calculation**

circle of the vehicle, or more) to free-cell status are marking the remainder as obstacles will achieve the desired obstacle growth automatically [see Figure 5(c). Furthermore, the absolute difference of the value of cells (other than those set as obstacle cells after the DT process) from the maximum value over all non-obstacle cells can replace the cell value as a risk of collision cost which can be incorporated into the path planning process. The local maxima of the DT field provides a digital version of a Voroni construction and can represent safe 'roadways' through obstacle space. A more complex DT example is shown in Figure 5(d), showing the global qualities of the methodology.

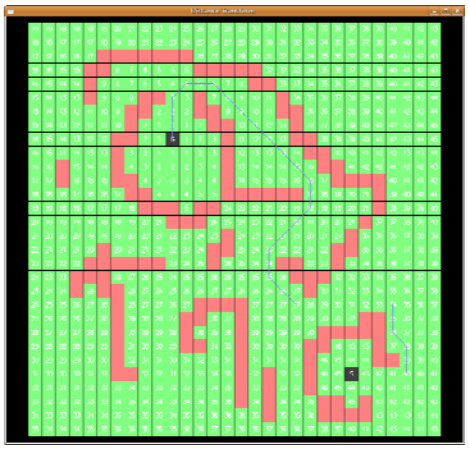


**Fig. 5(d) More Complex DT Example**

Two levels of path planning using the DT methodology are used in this project, one applied to the obstacle field data from the integrated Riegl scans one for the local Velodyne live obstacle field data. For the Riegl data a path from the current location to a nominated goal is calculated. As the position of the robot vehicle is changed a new path is calculated. Note that the DT need only be calculated only once for each new goal specification in this case. The goal point can be changed at any time, the DT being recalculated when required or simply continuously to avoid checking the goal change status. For the Velodyne obstacle field data case, the DT is always continuously recalculated (whether or not the goal status has changed),since dynamic obstacles may appear and , in any case, the robotic vehicle is moving.

For this project, given that the raw Velodyne 3D range data provides terrain height data, a roughness factor was be calculated at each free cell location based on the sum of absolute height differences from the candidate cell to each of its eight neighbours and this sum was weighted into the cost of entering a free cell, with 3:2 distance component included as well. All obstacles were grown by a nominated number of cells beforehand as described earlier.

## V. TELEOPERATIONAL NAVIGATION SYSTEM WITH FORCE FEEDBACK CONTROL

Figure 6 gives the block schematic for the whole teleoperational navigation system. The off-line Riegl data collection and localisation 'signature' data-base is entirely fixed and calculated prior to mission time. The environment Virtual Reality (3D plus colour) model [see Figure 7(a)] can be explored in detail at any time either before or during physical navigation. One may 'walk through' this virtual space at ground level or from any elevated viewpoint. During navigation one can either explore at will or use the localisation fixes provided by the system to position the viewpoint (elevation can also be changed independently). Live data from the Velodyne range scanner scan data, provided at 10 Hz rotation speed, is matched against the Riegl data-base (signature matching) to provide the current robotic vehicle pose. The local environment obstacle map derived from Velodyne range data is centred on this localisation fix with the vehicle direction of orientation always up on this map. The live 3D Velodyne data can be viewed simultaneously from a variable orientation view point and zoom.

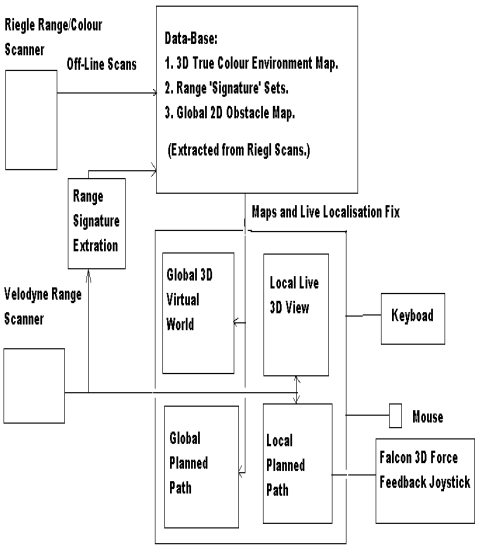



**Fig. 6. System Schematic**

A global environment obstacle map is also provided [see Figure 7(b)]. The global goal can be selected via a text file or using the computer mouse. The current localisation position defines the start point of optimal path trajectories to the goal (or least cost goal if there is more than one goal). The optimal (shortest) path shown on the global map is for grown static obstacle avoidance alone.

Fig. 7(a) 3D Virtual Reality Model

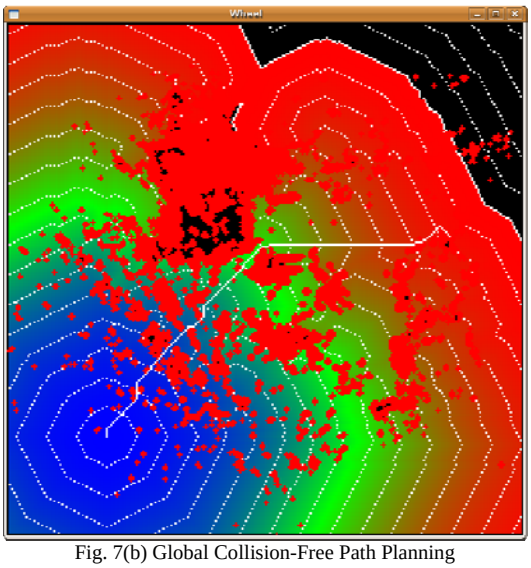
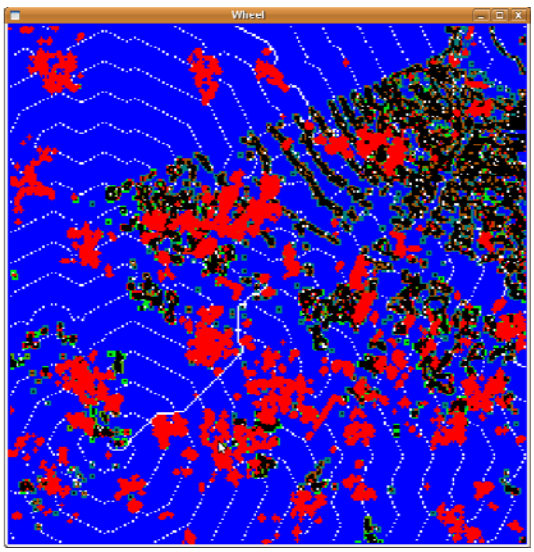


Fig. 7(b) Global Collision-Free Path Planning

The local obstacle/terrain roughness map shows live data updated from Velodyne data quite rapidly (e.g. at 0.5 second intervals). The local path trajectory (using a DT which accommodates distance as well as terrain roughness after growing obstacles a specified amount) is for advice to the operator with the centre of the map representing the current robotic vehicle location [see Figure 8] and the local goal selected using a computer mouse. In Figure 8, two different goal positions are selected; for the second image, it can be clearly seen how rough terrain is avoided at the cost of a longer path. The operator is free to choose a local goal consistent with the global path trajectory shown in the global display but can select any local position if variations to check environment details are preferred. It would even be possible (but has not yet been done) to make the local goal some number of steps forward along the globally determined optimal path as a default. Even when the local optimal path trajectory map reflects the operator's local goal selection the operator is free to ignore it. Given that the Velodyne 3D range data is live, any dynamic obstacle will be taken into account in the local path trajectory (but can not be so accommodated in the global fixed data unless the

Velodyne data is made to temporarily overwrite the Riegl data which has not been done, so far).

Now this is where the 3D Falcon force feedback joystick [see Figure 9] comes in. The horizontal/vertical movements of the joy stick control the driving of the robot vehicle (off the planned path if so desired) but force is applied to the joy stick to pull it back towards following the local planned path. However, each excursion away for the path defines the starting point for a new path so the force field is continually changing. Lightly holding the joystick allows the vehicle control to be consistent with the local planned path. Also, the third degree of freedom of the joystick (in our case in and out) is vibrated by a magnitude proportional to the terrain roughness factor calculated as described earlier so that driving over rougher terrain can certainly be felt by the operator. Only full field trials (not yet carried out) will determine how best to provide the force controls described above. It may prove necessary to provide some smoothing filters in the force feedback loop to reduce overshooting jerkiness. It would be hoped that the path preference and terrain roughness force feedbacks will give an intuitive feel to the operator and also effective navigation naturally without stress.



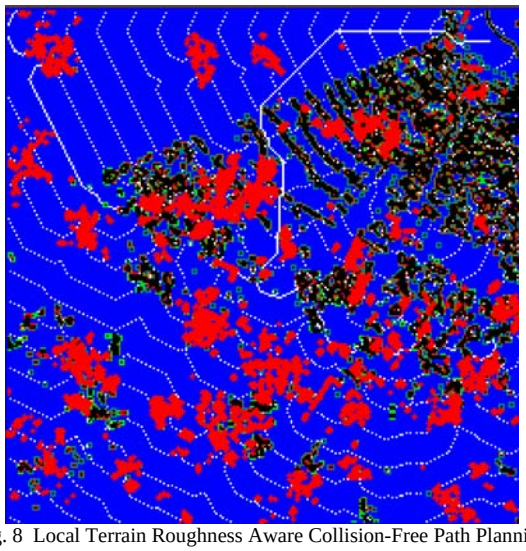



Fig. 8 Local Terrain Roughness Aware Collision-Free Path Planning

Fig. 9 Novint Falcon 3D Force-Feedback Joystick

## VI DISCUSSION AND FUTURE WORK

The three central elements of the work described here are as follows:

1. The Riegl range and image data is used to build a virtual world [14] of the robotic vehicle's work space and provides the 'signature' data-base for localising the vehicle by scan matching. The Virtual Reality world can be explored in detail at ground level or an elevated position either before or during navigation (in this latter case the current position and orientation

can be used for viewpoint determination if so wished).

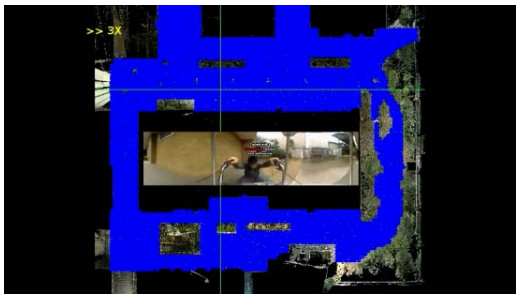
2. The Velodyne real-time 3D range not only provides 'signature' data for run-time localisation by scan matching against the Riegl 'signature' data-base but also provides dynamic local data on obstacles, ruts, moving objects, terrain roughness and the like whilst the robot is navigating and forms the basis of local path planning, force feedback navigation control and roughness vibration magnitude data in real-time.

3. The 3D Falcon force feedback joystick provides the operator with the capability of freely driving the robotic vehicle but with path planning guidance with preferred direction force feedback for driving and vibration feedback for terrain roughness monitoring.

Whilst not mentioned explicitly earlier in the paper, once the Riegl data has been collected and integrated, the vehicle can be confidently navigated at night since the virtual environment world is lit and the Velodyne data needs no ambient lighting to collect. Whether rain and/or smoke would seriously compromise this operation has not yet been investigated.

Also, in the future, more sophisticated data matching for localisation in very rough terrain might be explored to provide accurate and reliable 3D localisation fixes. Eventually, the navigation of smaller robotic vehicles in 3D man-made constructions may be possible using this approach.

In some earlier work [15], it was shown that an 'appearance-based' localisation method, where unwarped panoramic images collected on-board and compressed using Haar transformations were matched against visual signatures (similarly compressed) constructed off-line from the pre-collected range/image Riegl data base, could yield acceptable localisation results without using an on-board laser range finder. Particle filter methods were used to achieve approximate localisation. However, the accuracy achieved by this approach was not as good as that possible using range matching. The 'appearance-based' results would have been worse in more sparse environments where less position/pose discriminating views could be extracted. Furthermore the 'appearance-based' approach would be inoperable in poor ambient lighting conditions (or at night) whereas the range based system can operate in any lighting conditions. Figure 10 shows two snapshots of a localisation trace being calculated. The central inserted panel show the unwarped current view from the on-board panoramic camera. The test environment is a partially covered outdoor, paved, flat environment with high visual business. In Figure 10(a), the initial spread of the particles over the environment indicates a wide search to find the starting position by image matching. In Figure 10(b), tracking based on continuity constraints allows the particle scatter to shrink; the weighted average point is taken as the calculated location.




(a)


(b)

Fig. 10 Appearance Based Localisation Example

## VII CONCLUSIONS

This paper has introduced the idea of terrain roughness and path planning guidance for the teleoperational control of a robotic vehicle in out-

door rough terrain with force feedback to assist control and terrain roughness monitoring with the added advantage of exploring a virtual world of a 3D visual model of the working environment either before or during a navigation mission. Application areas such as bush fire fighting and search and rescue have been used to motivate this approach which allows both some degree of autonomous navigation to meld smoothly with teleoperational human guidance. Physical experiments to test the localisation methodology described in this paper have been successful in demonstrating the speed, accuracy and reliability of the approach, all of which were very satisfactory, even using the simple matching formulations described. More work on human factors need to be carried out to properly gauge the value of this approach to bridging the gap between pure teleoperation and fully autonomous navigation.

## REFERENCES

1. Jarvis, R. A., Terrain-Aware Path Guided Robot Teleoperation in Virtual and real Space, ACHI 2010, St. Maartins, Feb. 10-14,
2. Jarvis, R. A., A Go Where You are Looking Semi-Autonomous Rough Terrain Robotic Wheelchair, First International ICSC Congress on Autonomous Intelligent Systems, Deakin University, Geelong , Australia, 12-15 Feb. 2002.
3. Jarvis, R. A., Sensor Rich Teleoperation Mode Robotic Bush Fire Fighting, International Advanced Robotics Program/EURON WS RISE'2008,International Workshop on Robotics in Risky Interventions and Environmental Surveillance,7th to 8th Jan.,2008, Benicassim, Spain.
4. Leonard, J. J., and Durrant-Whyte, H. F. Simultaneous map building and localization for an autonomous mobile robot. In *IROS-91* (Osaka, Japan, 1991), pp. 1442- 1447.
5. Spero, D. (2007), "Simultaneous Localisation And Map building: the kidnapped way". PhD thesis. Monash University.
6. Jarvis, R. A., Very Rough Terrain Robotic Vehicle for Bush Fire Fighting Support, Proc. 36th International Symposium on Robots (ISR 2005), 29th Nov.- 1st Dec. 2005, Tokyo, Japan.
7. Jarvis, R. A., Virtual Reality Enhanced Excavator Teleoperation Proc. ISMCR'97 Workshop on Virtual Reality and Advanced Man-Machine Interfaces, Tampere, Finland, June 4-5, 1997, Proc. XIV IMEKO World Congress, Vol. IXB, pp.200-205.
8. Jarvis, R. A., Four Wheel Drive Boom Lift Robot for Bush Fire Fighting, 10th International Symposium on Experimental Robotics (ISER 2006), July 6-10, 2006, Rio de Janeiro, Brazil. Also in Experimental Robotics, Springer Tracts in Advanced Robotics 239,Khatib, Kumar, Rus (Eds.) ISBN 978-3-540-77456-3, 2008, Springer Verlag Berlin Heidelberg. pp.245-255.
9. Ray Jarvis and Nghia Ho, Robotic Cybernavigation in Natural Known Environments, Cyberworlds 2010 International Conference, 20-22 Oct. 2010, Singapore.
10. Lozano-Perez', T.: Spatial planning: A configuration space approach, IEEE Trans. Comput. C-32(2) (1983), 108–120.
11. LaValle, S. M. and Kuffner, J. J.. Rapidly-exploring random trees: Progress and prospects. In *Proceedings Workshop on the Algorithmic Foundations of Robotics*, 2000.
12. Jarvis, R. A., Robot Path Planning: Complexity, Flexibility and Application Scope, International Symposium on Practical Cognitive Agents and Robots, 27-28 Nov., 2006, University of Western Australia, Perth. pp 3-14.
13. Jarvis, R. A., On Distance Transform Based Collision-Free Path Planning for Robot Navigation in Known, Unknown and Time-Varying Environments, invited chapter for a book entitled 'Advanced Mobile Robots' edited by Professor Yuan F. Zang World Scientific Publishing Co. Pty. Ltd.1994, pp. 3-31.
14. Ho, Nghia and Jarvis, R. A. Large Scale 3D Environmental Modelling for Stereoscopic Walk-Through Visualisation, submitted to 3DTV Conference 2007, May 7-9, Kos Island, Greece.
15. Ho, Nghia. and Jarvis, R. A., Vision based Global localisation Using a 3D Environmental Model Created by a Laser Range Scanner, Proc. IROS 2008, Nice, France, Sept. 22-26, 2008, pp. 2964-2969.

# Development of a Rich Picture editor: a user-centered approach.

Andrea Valente
Dept. Electronic Systems, Aalborg University EIT
Esbjerg, Denmark
av@es.aau.dk

Emanuela Marchetti
Warwick Business School, The University of Warwick
Coventry, UK
emanuela.marchetti.10@mail.wbs.ac.uk

***Abstract***— This paper describes the development of a software tool to support rich pictures creation for object-oriented analysis. This software is useful both as an e-learning tool for bachelor-level students, as well as for practitioners working with agile methodologies. The transposition of manual rich picture practice into software proved difficult, therefore, we decided to follow a user-centered approach: design and implement a prototype with basic functionalities, then run a usability test with a few students and professionals. The feedback collected in the test validated our hypothesis circa the need of software support for the authoring of rich pictures, but also forced us to re-consider the design of our prototype. To gain a deeper understanding of the students' working practice, we also reviewed rich pictures from past student projects. All the information gathered through our study is guiding us in the design of the tool next version. At a more general level we realized that modern object-oriented development methodologies, such as agile methods, are informed by design, hence they sometimes assume design skills that programmers do not have or do not value.

***Keywords- rich pictures; knowledge acquisition; object-oriented analysis; qualitative tests; learning***

## I. Introduction

Rich pictures [1][2] are more and more part of object-oriented analysis and design courses (OOA and OOD courses). At our university, bachelor students in Computer Science as well as Engineers are required to perform analysis in small groups (3 to 6 members) and draw rich pictures as part of their project documentation [3]. Usually rich pictures are created with low-tech support, such as whiteboards or pen and paper. Students sometimes adopt some general purpose software, like a painter or a diagram-drawing tool.

Rich pictures represent knowledge about a domain (similarly to Novak's concept maps [4]), and should guide the developers during the definition and construction of the system's early prototypes. However, using a generic tool instead of a specific one has known disadvantages (see [5]). In the case of OOA it means that fundamental concepts are missing and that the knowledge acquired is not immediately re-usable, especially for generative purposes. Hence, it is not possible for an analyst using a generic tool to *translate* rich pictures into rough software prototypes of the system under study. It would of course be possible to use one of the many formal-methods software tools, but they require training from the part of the students, and mostly work with rather complete and detailed knowledge of a system, being therefore typically unusable in the analysis phase or when acquiring knowledge incrementally.

Considering all this, we decided to develop a software tool specific for the creation of rich pictures, to be used in OOA. This software should be useful both as an e-learning tool for bachelor-level students learning OOA and OOD, as well as for practitioners, working in small teams adopting agile development methodologies.

However, transposing the manual rich picture practice into a software tool proved difficult, so we decided to follow a *user-centered* approach and involve students in a usability test. The feedback collected during the test greatly eased the task of defining the main features of our tool.

In the following section we present an early version of our tool and discuss our ideas, sources of inspiration, and related works. Section III explains how the usability test was constructed and run, and what we discovered observing our students interacting with the tool and later interviewing them. In Sections IV and V we discuss the test and how the feedback from the students is guiding the next iteration of the tool development. The new version of the tool, with a new GUI and extended features, is outlined in Section VI. Section VII concludes the paper.

## II. Software support for rich pictures

According to [1] a rich picture provides "a broad, high-grained view of the problem situation", and it shows *structures*, *processes* and *concerns* (or *issues*). It is also remarked that there is no best way to construct a rich picture. From this consideration we derive a requirement for our software tool: it should not impose a specific work-flow to its users.

When rich pictures are used for OOA, structures become visual representations of objects or grouping of objects, while processes are understood as events, changing the state of one or more objects instantaneously (as explained in [3]). As for concerns, they are often simply notes written in natural language aside of the different objects in the picture. Our tool should therefore be a drawing program, and it should allow users to create frames (to visually represent objects), eventually nesting them, to group many frames together into one. Furthermore, users should be able to describe events involving many frames, i.e., specify the processes at work in the system model. It should also be possible to write natural language notes, to support concern identification. We want our software tool to help the user to

explicit the knowledge captured by one or more rich pictures. This will provide support for an automatic generation of (skeletons of) executable prototypes.

*A.Related tools*

To our knowledge there is no software support specific for rich pictures, so we decided to proceed on two fronts: first we surveyed existing software tools that could generally relate to visual editing and conceptual modeling [6], and at the same time we established our own requirements for an authoring tool specific to RP, and to be used in object-oriented development.

The survey covered concept maps [4] and text graphs [7]. Concept maps have a very established community, a clear definition and many good software tools. They have been used for many decades in fields like knowledge acquisition, e-learning and knowledge visualization. A concept map is typically a graph structure, constructed from labels containing natural language phrases, and arrows linking labels together. The focus is on the definition of concepts, type-like entities, while rich pictures show more concrete, instantiated examples of a system's state and dynamics. Text graphs are an interesting attempt at making concept maps meaning more precise. However, they are text-oriented and they offer no clear way to represent different steps in the evolution of a series of concepts. While text graphs are not developed with rich pictures in mind, they suggested a direction of inquiry: what happens when text is replaced with pictures, in a text graphs? We explored possible answers to this question in [8], where we also discuss criteria for conceptual modeling software tools.

Another option for us was to adapt existing visual editors to RP [9], therefore we experimented with a few products as well as discussed the matter with our student testers (who have also independently tried to author their rich pictures with available software). The most interesting tools we considered are Visual Paradigm for UML [10], Microsoft Visio [11] and Dia [12], and Visual Knowledge Builder [13].

Visual paradigm for UML [10] is a specialized tools for UML-related development activities, such as design of state machines, use cases, class diagrams, and deployment diagrams. In the user guide, visual paradigm is defined as: *"a powerful, cross-platform and yet the most easy-to-use visual UML modeling and CASE tool."* A very comprehensive tool, as other modern CASE programs, it can import an existing object-oriented program and automatically generate diagrams from the code. These tools are very good and integrate well many diagrams into a coherent detailed specification of a system. Systems can be defined incrementally, but the notation is built-in and standard (usually from the family of the UML diagrams). Visually appealing, visual paradigm provides a friendly and innovative GUI. However, its goal is not support knowledge acquisition: if a system is yet to be defined, what is the point of keeping strict relationships between its various sub-components and views. We are more interested in suspending validation and letting developers *explore* and correct their diagrams through discussion.

Both Microsoft Visio [11] and Dia [12] are diagram editors; the first is proprietary, while the second is a GTK-based GNU tool and is often introduced as a free alternative to Visio. We analyzed Dia in greater detail and found it a good visual editor for diagrams, with many predefined *shape packages* (e.g., for UML diagrams, electronic circuits as well as various business diagrams). Dia has a palette and a drawing space, and users work by dragging shapes from the palette into the drawing space; then they can customize properties of the shapes and connect them by means of various types of connectors. Interestingly, the set of libraries can be extended, as new shapes can be described by XML files. It is also possible to design custom shapes directly in Dia, and the custom shapes can also be given special attributes. It is clearly possible to use Dia for RP, but being a generic tool, the burden of interpreting the diagrams as RP will reside solely on the users. As discussed in [5], it is not always the best choice to adopt general purpose tools for specific practices (as also emerged from our test, detailed in Section IV).

Since working with RP requires *spatial reasoning*, it is relevant to consider software like the Visual Knowledge Builder (VKB) [13]. It uses *incremental formalization* to simplify the expensive and time-consuming task of defining knowledge. Many of the goals of VKB are strikingly similar to ours. VKB is visual, but the graphic elements at disposal are simple geometric shapes, little freedom of expression is left to the author. VKB allows users to proceed incrementally from concrete examples of structures, towards more general patterns, type-like in nature. However, VKB seems to be more oriented towards analysis than synthesis, and it bears little relations with object-orientation and OOA.

Our general conclusion is that these tools fall into 2 opposite categories: they are in fact either *too specialized* (e.g., they work very well with a subset of UML diagrams), or *too general*. What we would like to achieve is a tool in between Visual paradigm and Dia, and that can adequately represent the concepts required for RP editing. This is why we decided to design and implement our own RP software.

## III. The early prototype: FSSE 2009

The new tool is called *Free Sketch for Software Engineering 2009* (called FSSE in the rest of the paper) [9]. The GUI of our tool is visible in Figure 1A. It is composed of 2 windows: the largest one is the main drawing area, where users draw their rich picture, and a smaller window called *palette* that contains type-level information about the elements drawn in the rich picture.

The typical work-flow of a user creating a rich picture in FSSE would be:

- Create a new, empty FSSE project.
- Draw an image in the background of the main window (using an external painter program) or alternatively import a scanned hand-drawn image. This background image serves as initial draft of the rich picture (see Figure 1A).

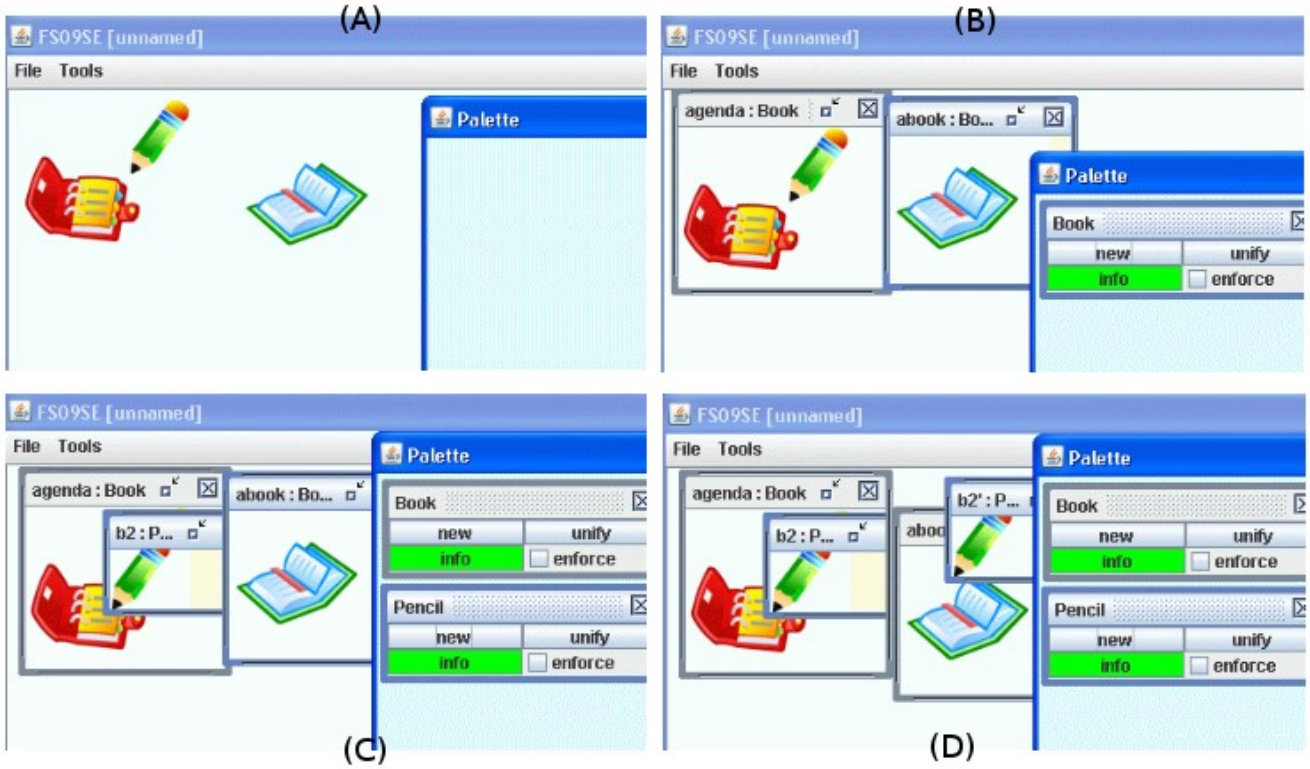



Figure 1: The GUI of FSSE. In the top-left part of the figure (A) the user imported a background image, representing some objects of her rich picture. The second part of the figure (B) shows how the user can convert background images into frames, with names and tags; in (B) "agenda" and "aBook" are frames, tagged with tag "Book". The tag "Book" is also represented in the palette (on the right). In part (C), the image of the pencil is converted into a frame named "b2", then nested into the "agenda" frame; the tag "Pencil" is now represented in the palette. The last part, on the bottom-right of the figure (D) shows how the user can use the "Pencil" tag to create a new "Pencil" frame, then place it close to "aBook".

- Select rectangles out of the background image. Each selection turns into a frame, that the user can move around and clone, to obtain multiple copies of the same frame.
- Each frame can be given a name and a list of tags. Names do not need to be unique, and tags are like types. Tags in FSSE are a clustering device, like tags in blogs.
- More and more frames will be defined, so that the initial background image will be reconstructed by frames. This structuring process starts from a flat image, and converts it into a rich pictures made of objects, i.e., frames (as in Figure 1B and 1C).
- Frames can have internal details; to declare that a user simply selects a rectangular area inside a frame, and a new frame will appear, nested in the selected one. It is also possible to insert a frame into another one, via *drag-and-drop*.
- The palette window is automatically populated, and contains at any given time a list of all tags used in every frame in the main windows (without repetition). This incremental creation of tags in the palette is visible in Figure 1B and 1C.
- A tag in the palette (see Figure 1D) can be used to create a new frame, instance of that tag. Each tag also provides information about the relationships between itself and the other tags, such as cardinality and optionality of associations.

Our tool does not force users to decide in which order to perform their structuring of a rich picture. For example the division of the initial background image into frames can be mixed with the declaration of the internal structure of the frames.

Users can even decide not to assign names or tags to their frames. A frame without names nor tags could be used to group other frames. This means that frames do not correspond exactly to the objects in an OOA. Frames are in fact more un-structured than objects, and become representations of objects only when users decide to assign names and tags to them.

To implement frames we drew inspiration from *mobile ambients* [14]. Dynamic tree-like structures with names and types, ambients can easily model objects and proved a good metaphor in the design and construction of FSSE.

In our tool, a frame can have multiple tags, which corresponds to an object with multiple types (or classes). We

designed FSSE to allow for multiple hierarchy, in this way a rich picture could have rich and/or loose relationships among tags, and the user can decide, at a later time, to clean up her tags into a single inheritance tag system. This kind of alteration of tags relationships (i.e., relationships among classes) reminds of refactoring practices.

As soon as a tag is used for a frame, FSSE automatically adds it to the palette window. Moreover, our program analyzes the relationships among tags, and finds out the *typical structure* of a tag. According to what is depicted in Figure 1D, "agenda" is a "Book" and contains a "pencil", that is tagged "Pencil". However, the frame "aBook" is also tagged "Book", but it does not contain any internal frame. Therefore, FSSE will describe the "Book"-tag as having an association *0 to 1* with the "Pencil"-tag.

Events are not yet supported in FSSE. It was unclear to us, before running the usability test with our students, how to best add them. Concerns are not present either, but they can be expressed by writing comments directly on the background image of the rich picture.

## IV. Tests

### *A. A qualitative usability test: set up and task*

At the current development stage of FSSE, a preliminary usability test was needed in order to complete or even to change the tool radically. This test is based on our hypothesis that students may find RP more relevant and useful to their project work, if they could edit them on a specific software tool. Such tool should also allow them to re-use RP for generative purposes, turning RP into an integral part of OOA.

Participants to our test were a professional programmer and four engineering students at the 5th semester of their bachelor, who have recently started a course about OOA and OOD. Our aim was to evaluate how users may perceive a tool like FSSE, if it is seen as useful, easy to use, and if it adequately supports work-flow, for individuals and groups. The students were divided into two groups and were invited into a classroom, one group after the other. The students were sitting at a desk, with a laptop running FSSE, and we were in front of them, observing their reactions, taking notes and filming them with a video-camera. The laptop was connected to a projector, so that we could see (and film) their actions on the wall behind them (Figures 2A and 2B).

The test was articulated into four stages: first we showed the students a 5 minutes video-tutorial, then we introduced them to a task, and we left them free to familiarize with the tool before starting; at this point we started filming. The task was similar to the one shown in the tutorial, they had to create one or more rich pictures, identifying objects, classes and events, regarding a pizza restaurant (see Figure 3). A customer can order a pizza from a menu talking to a waiter, the pizzas have to be baked and can be served with wine or other beverages. Finally the customer pays the waiter and a conflict may emerge between them about the order.

After the task completion, we asked them a few open questions about their impressions of the tool. A list of questions was prepared, but it was intended mostly as a reference.

- How did you like the tool? General impressions.
- Given you experience with object-oriented modeling, do you think the tool can facilitates object-oriented analysis and design or no? How and what will you change?
- Do you think that the tool makes object-oriented analysis and design more understandable for users or not? How and what will you change?
- How do you think it will be possible to define events in Free Sketch, within the current user's interface and how could it work?
- Do you think you would like in future to use a tool like this in your work or not? Why?

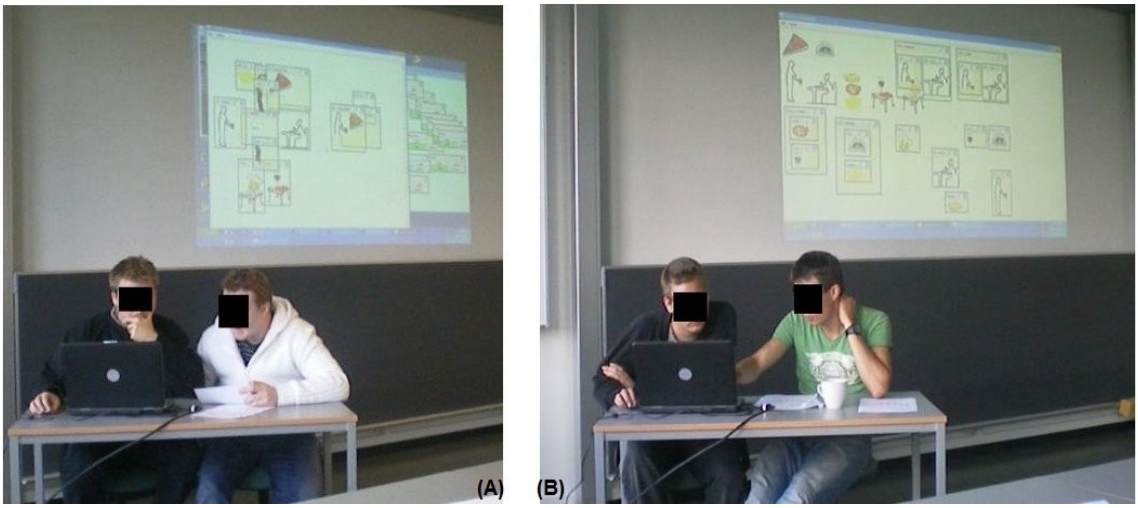



Figure 2: Two groups of students (on the left and on the right) trying to model events in FSSE. Since events are not actually part of the features of FSSE, each group freely invented a way to express them: the result was a couple of different approaches. The first group (on the left) modeled events by clustering of frames and arrows. The second (right) nested the frames involved in the event in a new frame, representing the event itself.

- How do you think the tool supported flow of team work? Did it facilitate team work or made it more complex? How could the tool be improved?
- Other comments? What other changes will you suggest to make the tool more effective in supporting object-oriented analysis and design in software development or its understanding from a student's perspective?

During the test in fact we started from the first question and then we adapted to the students' comments, who sometimes covered several issues at one time or even proposed new issues. For practical reasons we could not meet the programmer in person, we gave him the program and the tutorial, he solved the task in the tutorial and sent us feedback by e-mail.

In designing our test we referred to user-centered qualitative approaches, like ethnographic observations and analysis of video recordings [15][16]. Our aim was to gain a detailed account from users about their working habits, their experience of the tool, how they would like to work and eventually be supported by a tool like ours. These data were intended to be used in a new development iteration.

The task was designed as a typical modeling problem, of the kind they already faced during their OOA and OOD course, so that they could reflect upon their own experience to evaluate the tool. It was also our interest to observe how FSSE fitted within the team work-flow and how it affected *reflection in action*, intended as a process of critical thinking while performing a skilled practice [17].

Concerning the questions, we referred to the method of situated interviews [16], that prescribes to interview users in their context of practice, starting with open questions and gradually focusing on the details of users' statements and ask for examples. We preferred interviews to questionnaires to find out what really mattered to the students and to show them that we cared for their contribution, and this was explicitly appreciated by one of them.

### *B.Collected Data*

The students responded quite positively to the test and the prototype, it seemed as we were on the right track. They were relaxed with their mates, probably because they were already working together in the same group for the course and the semester project. They sat one aside of the other, one interacted with the computer, the other read from the paper with the task description and often pointed at the screen with one finger, then they talked a lot deciding together on what to do.

We expected the time required for the test to be around half an hour for each group, but in fact it took one hour, as they used extra time to get familiar with the interface. However, they all said that the purpose and the interface of the tool were easy to understand.

Surprisingly for us, drawing appeared as a main concern to all the testers, they felt visibly uncomfortable when they needed to draw new icons, specifically arrows and the menu for the restaurant. The first group expressed their uneasiness exchanging a worried, ironic look, then after several attempts they drew a menu and arrows to connect the pizzas to it (as visible on the back of in Figure 2A). A member of the second group said ironically: "Ok, we suck at drawing!", then they modeled the menu as a new frame with the pizzas nested inside, avoiding to draw.

The feedback we received from the programmer was very similar, he wrote that he likes the tool, and he also remarked that he does "not want to play with graphics, it sucks!", when analyzing a system. He then suggested to add a library of free, pre-drawn icons and arrows. In this way he

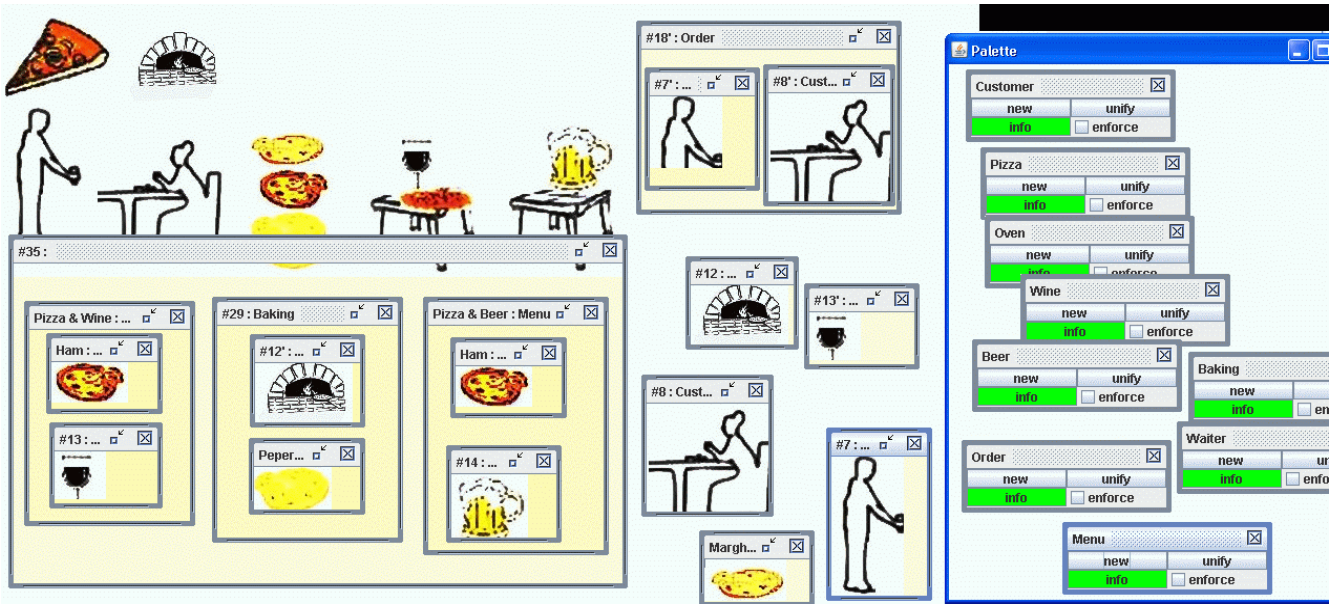



Figure 3: The "pizzeria" task modeled by one of the students groups, using FSSE 2009.

proposed a constructive solution to the same problem that was signaled also by the two groups.

These reactions revealed programmers' perspective on agile methodologies, which include soft skills, such as prototyping and drawing to make rich pictures and storyboards. These skills are taken from the field of design, therefore do not belong to the curriculum of a computer scientist or an engineer, and are not even part of their system of values.

Through the interviews we realized that drawing on paper is perceived as an annoying interruption in the process of reflection in action. According to them, it takes time to make a decent icon, approved by the whole group, as they have often "to draw, erase and draw it again", hence "just having a tool would help!". Moreover during the test they were quite precise in selecting icons and spent time erasing the superfluous parts in the external painter, to make them more *readable*.

Their quotes and actions show that, despite their dislike for drawing they want nice icons in their rich pictures, but do not want to do them by themselves. In this sense, features like automatic insertion of pre-made icons or creation of icons through selection from background pictures (as currently available in FSSE), do provide a smoother work-flow also from a team work perspective. It was also proposed, both from students and researchers, the possibility to introduce collaborative user interfaces, to turn the main drawing window of FSSE into a sort of shared, remote desktop.

Definition of events is central during OOA, but events were missing in the prototype tool that we tested. Nevertheless, the task assigned to the students required to try and represent events. We wanted to see how the students might interpret events representation within the given FSSE interface. They all expressed their perplexity for the lack of support, but found their own way to solve the problem. Interestingly they all tended to represent events as scenes of a storyboard, but they kept the approach they used to define complex objects. The first students grouped a few frames and connected them with arrows (Figure 2A), while the others grouped frames by nesting them into a fresh new frame (Figure 2B).

Finally the students seemed to find confusing the distinction between names and tags, so that they discussed with each other how to use the two labels to keep their rich picture coherent. However, it did not take long before they understood that tags work as types and names are just arbitrary identifiers to be assigned to the frames. One of the students showed to be a little frustrated by this ambiguity and said: "if it is a type, why do not call it type!". In FSSE we wanted to use the term *tag*, since tags are supposed to be used with more freedom than types (see Section II).

Moreover, to facilitate overview of the system created, a student proposed that when a frame is selected, it should be highlighted, together with the other frames sharing its tag.

Furthermore, FSSE was appreciated for its flexibility, enabling users to keep their favorite work-flow and their understanding of rich pictures making. Such flexibility implies that users can start modeling from a chosen level of abstraction, and mix the various activities as they like. This is what is called *middle-out modeling* in [8].

One of the students, who tried a few generic software tools in RP editing, commented: "the nice thing is that this tool doesn't impose me a specific way of thinking, it doesn't assume I am stupid!". Hence we realized that work-flow flexibility can give a feeling of not being patronized, by providing users more *control* on their work.

### *C.Theoretical framework for usability test*

Our usability test was conceived to actively involve the students in the design process, in a simple way. It is based on *User Centered Design* qualitative research principles [16] [18]. A prototype was provided to them and they were asked to solve a simple modeling task, simulating their everyday work practice augmented with our tool. The prototype was a working software, yet it was a mock-up as did not have all features implemented. Specifically no support for events was provided, so that the students could inspire us about how to design this particular feature, which appeared to be quite difficult. Therefore, our prototype did not support all the actions required by the task, providing only a rough feeling about how they might be supported by the finished product.

We expected that when the students realized that a specific feature or a standard way to represent events were not given, they would have shown a feeling of perplexity, but found their own way to do it, bringing new ideas to the design process.

Our approach involves principles similar to the ones discussed by Suchman [20][21]. She points out that to be able to reconstruct artifacts as objects of investigation it is necessary to alienate them, so to be rediscussed and understood in action, with the active involvement of users. In this case, we distanced ourselves from our program, by neglecting its completion, so that we could re-conceptualize it together with the students. We willingly introduced an incompleteness, which worked as a kind of provocation to the students, creating a bit of frustration. As expected the students were able to get over their initial uneasiness and to affiliate with the program, deciding on one important feature. In this way the program was designed as close as possible to the context of use, with users expressing their point of view about new possible versions. Some of them showed appreciation for being invited to the test, as they realized that we actually wanted to share with them our affiliation with FSSE, when it was still in the beginning of development.

Another aspect that was fundamental at that stage and required involvement of groups of students, was to evaluate the impact of FSSE on team work. The test and the analysis of RP in fact showed that the students prefer to work with a software tool, for several reasons, including their dislike for hand-drawing. But as the activity of sketching on paper fits well team work, as it can be done by more individuals operating on single paper sheet, the same thing is not obvious regarding a software running on a computer. The computer itself has an affordance to support one individual operating and this was clearly visible during the experiment.

The students participated at the test two at a time, and already like this we saw that one student worked at the computer, directly using the tool. The other student instead sat on one side and looked at the paper with the task description, but they both participated in decision making (as in Figure 2A and 2B). There was no strong reaction about this interaction style from the students' part. It is possible that they did not feel disturbed as the set-up suggested two different roles to be chosen within the pair, or simply because they are used to this kind of dynamics from their everyday practice of software development. However, it is our intention to run a user study with a new version of the tool in the fall semester and observe students in the act of analyzing their problem in groups. We expect that this study will allow us to see the program in action, evaluate our findings from the preliminary test, and to identify forms of emergent interactions that might facilitate group interaction in RP editing. These new data will be analyzed in order to improve the program and make designerly activities, such as OOA and RP editing, more engaging and meaningful from the perspective of technical students. In our view, this aim will be achieved re-situating RP creation, now perceived as an independent pedagogical activity, within software development, so to be perceived as an integral part of it and not as a superfluous exercise.

## V. RE-CONCEPTUALIZATION

### *A. Analysis of RP across past reports*

Reflecting on OOA&D courses through the past years, we had the impression that students generally fail to recognize the importance of RP in the development of a software, and certainly do not like to make them. Generally it seems as they consider RP as compulsory project documentation, explicitly required by the teachers, but not particularly meaningful for development, which is considered by our students the most relevant part of the project.

In order to investigate further our impressions, we analyzed a few students' project reports containing RP (or sometimes loose re-interpretations of RP), to see how students actually related to the rich pictures as a tool, and as part of OOA&D.

We collected 11 reports written through the past seven years: 7 of them were intended for a bachelor-level OOA&D course, for which RP are a specific requirement. The other 4 were instead intended for more advanced courses involving software development (for example a master-level course in computer games and interactive systems), for which RP are not mandatory, as the students are supposed to choose independently their method. All the 7 reports intended for the OOA&D course contain RP, 4 of them even provide a definition of RP. Instead only one report out of the four intended for more advanced courses has a RP. Hence it seems as RP are made only when explicitly required, in fact it was interesting to notice that in some cases the same group of students made a good RP for the OOA&D course but did not make any for more advanced courses.

Interestingly all analyzed RP make use of explicative texts to clarify the situation described. Furthermore, the textbook for the object-oriented course [3] recommends to make a few RP during the system choice phase, as a way to generate discussion and facilitate requirements definition for the system under development, and some teachers also suggest to proceed like this in class. Despite all this, only one out of the 7 OOA&D reports has 2 RP representing the same situation from a different focus; all the other only contain 1 RP.

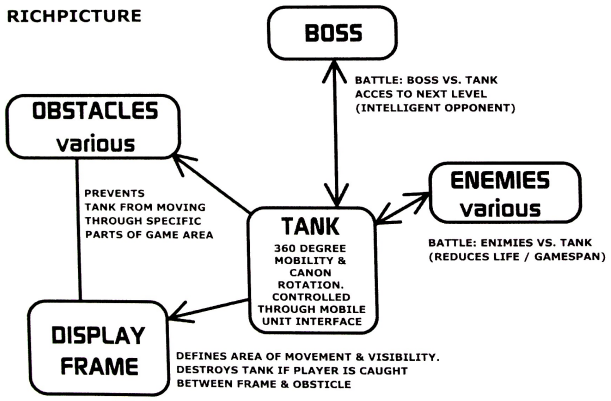



Figure 4: Rich Picture from an advanced course. In the report it is said that it was edited as a support for the reader, not for analysis.

The diagrams provided in the other four reports (including the RP) might resemble RP, but they mostly describe use-cases or state diagrams, showing once more the focus of our students on the technical aspects of system development. Interestingly the only provided RP, visible in Figure 4, is used in a quite improper way. The students wrote that it was drawn to "show the problem domain and possible conflicts to the readers after all decisions were made". This seems to confirm our impression that the students consider RP as a tool for readers (teachers of stakeholders), but not to support analysis as they are supposed to. In their RP, users and context of use are not represented, and conflicts are missing too. Representation is based mainly on written text, probably because of their general dislike for drawing.

Furthermore, considering the representational details of the RP we could see that only 3 RP are handmade, all the others are instead edited on a computer tool. The students follow different approaches in representing the visual structure of RP: some follow a sequential structure while others prefer a circular representation, at which center is the system to be developed, the context of use or the users.

Only four reports include two RP, one for the current situation and, in opposition, another for the new improved one.

Finally, conflicts seem to be a bit neglected; only 4 reports out of all 11 show conflicts. One of them, represented in Figure 5, has only the "tradition versus change" meta-conflict, as given by typical examples in making RP [1][3].

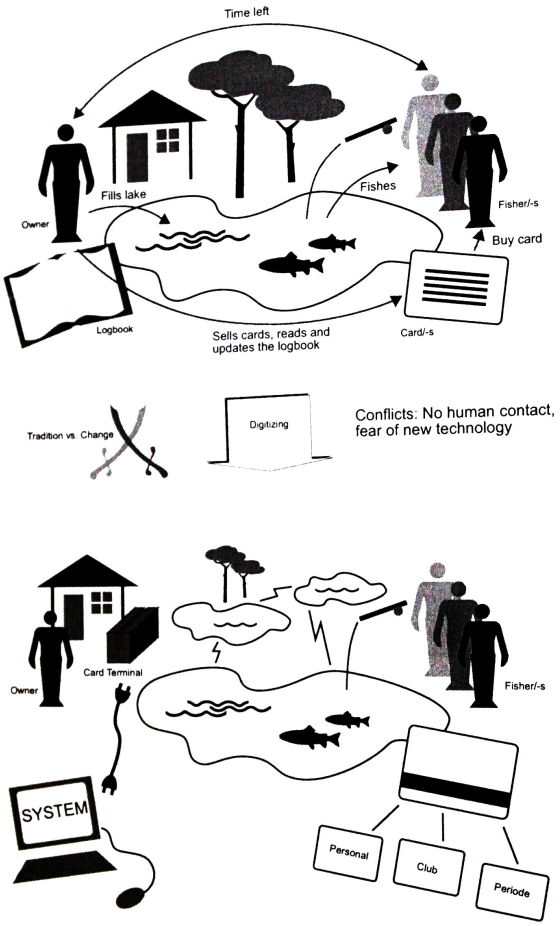



Figure 5: Typical example of rich picture from a students' project. It is edited through a software tool and represents only the "Tradition versus Change" conflict.

*B.Discussion*

Analysis of students' reports shows that students abandon RP as soon as they go further with their studies, cutting them out of their work practice. This phenomenon could be related to the fact that students underestimate or did not understand the importance of requirements gathering and analysis, preferring to get to the technical part. It might also be that they underestimate the use of sketching, still giving importance to knowledge acquisition. A possible reason could be that the tasks they receive for the projects are either too technology oriented or too simple to require a deep analysis. This certainly has to do with the fact that it happens quite seldom that the students receive tasks from potential clients/users from the real world. In most cases it is the teacher who defines such problems and assigns them to the students (in contrast with the "complex and messy problematic situations" discussed in [2]).

Hence these problems might be too defined from the beginning, so that it is immediately visible how to capture the elements of the problem domain in terms of object-oriented features like classes and methods. In this way RP making becomes a superfluous exercise, just to show the teachers that the prescribed path has been followed correctly.

Assuming this explanation as correct, it means that knowledge acquisition and analysis in software development are in fact designerly activities, in the terms expressed by Rittel and Webber [22]. They define design as a process aimed at framing and solving "wicked problems", in which the presence of messing factors, such as people and social interaction, makes it impossible to find easy or optimal solutions. However, still in more recent studies [2], problem framing is recognized as a valid tool, and it automatically embodies suggestions towards suitable solutions. Therefore, if the problems assigned to our students are not *wicked* enough, and they did not require any framing, then the technical part is the only one left for investigation.

The same reason could also explain why conflicts are often left out of RP. If the problems are too easy, it might even be necessary for students to artificially invent possible conflicts for their RP. Conflicts should instead emerge through analysis of a messy situation, taken from the real world, that demands technology supported solutions.

Exploring more this angle, it is no surprise that RP were originated within Soft Systems Methodology [1][2]. They are related to situation boards designers use to represent users' dilemmas and context of use. Situation boards provide a support for reflection and design, intended as a creative and exploratory process, within the design team eventually involving users too [15][18]. RP should be used in a similar way, therefore, they can be defined as a designerly tool. According to Stolterman's definition [19], there are no step by step instructions about how to make RP, as they are supposed to be flexibly adapted to the situation to be represented and preferences of individuals or the group. In fact we saw that the students are actually keen on structuring RP in different ways. But we could also notice in past years that this flexibility may generate confusion: there is no specific way to make a RP, yet students can still make them *wrong*, not rich enough or missing key elements needed when later modeling the system. Another source of confusion is the fact that rich pictures, use cases, and state diagrams all contain some of the same pieces (such as users or events). In face, some diagrams from the four reports intended for the game course tend to mix internal details of the system to be developed, with context of use, and conflicts about the application of the system.

In conclusion these issues may be solved if students received their tasks from actual clients, like for example a company. If that was not possible, the teachers could make the effort to provide messy problems, maybe taken from news papers or other *real-world informed* materials. Hence students could be provided with heterogeneous stories describing the same problem from different perspectives (e.g., discussions about the different ways to administer existing power plants and renewable energy sources). At this point the students would be forced to analyze such material, to frame the general problem, isolate one or a few specific issues to focus on, identify core elements, actors, events, and potential conflicts in the original and in the new changed

situation. Hence RP might gain recognition as a useful tool that allows developers to find a focus in the messy real-world and explore more before committing to a particular system definition.

### C.Re-conceptualization of RP as knowledge acquisition

Reflecting on the results gained from the preliminary test and the analysis of RP in past project reports, we identified a typical work practice related to knowledge acquisition and pre-analysis, which are the initial phases of software development, and RP editing. This work practice is what our software tool should facilitate, when finished.

RP creation is a preliminary design activity, the stage where developers must frame a messy problem in order to find adequate solutions, focused on object-oriented technology. The RP creation process is quite complex, and it is definitely a form of *reflection-in-action* as defined by Schön [17] regarding design and planning. In this practice experience and improvisation are deeply intertwined, as expressed by Ingold and Hallam [23]. Moreover, it is a social practice, since decisions must be taken by a group of developers.

Schön, in his book "The Reflective Practitioner" [17] provides a deep analysis of professional practice, reconstructing how professionals act in their everyday work and reflecting on implications for education. In our case we are dealing with bachelor students from technical departments (Computer Science, Engineering, Medialogy), who have to learn object-oriented analysis and design in their curricula. During their course the students are supposed to learn theory and practice of object-oriented software development, usually by working at a mini-project that spans the duration of the course. Moreover, the students are typically developing their semester projects at the same time as they attend the OOA&D course, and can decide to apply some of the concepts learned to the larger semester projects as well.

As discussed by Schön, the students are supposed to acquire a repertoire of examples ([17] p. 138) regarding application of techniques, theories and practical knowledge, based on their project experience, to support their future working practice. Working at their mini-projects, students are training in analyzing the given problems and in applying the knowledge they gained through lectures and text books, in order to develop technology supported solutions. This kind of practice is called by Schön *reflection-in-action,* and it is defined as a reflective conversation with the material of the design situation ([17] p. 165). Sketches, like RP, represent virtual worlds through which the practitioner can make exploratory experiments, to investigate possible solutions for her task. New decisions will be taken, reflecting on technical and social implications through these exploratory experiments, which *talk back* to the developer.

Moreover, RP creation is also a social process, since all group members are supposed to participate. In this sense it involves an improvisational component, as defined by Ingold and Hallam [23]. Improvisation is a *relational generative* process, it is functional to the creation of new culture and implies that all actors are responsive to each other and the context. It is also *temporal* as it embodies a certain duration, that is being defined by an organic sequence of actions articulated through time [23]. All these aspects are present in RP editing, which unfolds as a participatory knowledge acquisition, leading to the identify objects, users and dynamics of the system to be developed.

Considering all this, the software tool we are developing must be re-conceptualized, to support reflection in action within a social context. Thus, as already mentioned, FSSE should be a designerly tool that does not impose a step by step guided practice, yet it must have a specific affordance for RP editing.

Furthermore, FSSE should allow developers to structure their own elements (such as objects and events) when editing one RP. In this way developers should be able to create a sort of *kit of tools*, that is supposed to speed up the process of editing future RP too. In more general terms, developers should be supported in creating a rough *visual domain specific language.* Therefore, in designing FSSE, balance between specificity and openness represents a fundamental dilemma.

## VI.Next iteration: FSSE10

Considering the details analyzed in the RP we can deduce possible features for the new version of the program. First of all we noticed that only a few RP were handmade, this confirms our findings from the test that technical students dislike to draw and prefer to use a graphical software tools for their RP. This behavior is compatible with our hypothesis that students consider RP as something required by teachers, and if edited at the computer, they look better in their reports and are more readable. However, even when created with software tools, RP are clearly structured in a personal way, independently from the tool used.

In terms of designing our tool this implies that we have to allow students to freely choose their representation style, a principle that fits within the definition of a designerly tool [19]. Refining FSSE to be a better designerly tool for RP is our main goal for the next iteration; the new version of the tool will be called FSSE10, since it will be finished and tested in 2010. From a functional point of view, FSSE10 needs to provide better support for the 3 central elements of RP: structure, processes and concerns, and possibly present a simpler and clearer graphical user interface (GUI). In the next sections we will discuss the design of FSSE10.

### A.Streamlined GUI and new palette

Considering our observations circa the way students work with RP and with FSSE, we think nesting of frames complicates the GUI; therefore nesting will be replaced by stacks of re-positionable notes (a concept similar to *piles* in the BumpTop virtual desktop [24]). The new metaphor should be that when a frame B is stacked on top of another frame A, then B is inside A, or B part-of A.

Moreover, the new GUI will integrate free-hand drawing: to draw we currently rely on a free external painter (Java Image Editor, by JH Labs). Internal painting capabilities will

provide a more uniform environment and improve the flow when drawing rich pictures.

Many students seem to like to add explicative comments to the RP or to single elements of it. This practice, related to RP concerns, will be supported by allowing them to place text bubbles in the rich picture.

The palette is also undergoing significant changes: it will look much more like a simplified UML class diagram. The terminology used in FSSE10 will therefore be more in-line with object-oriented jargon. *Tags* will be called *classes* and *frames* will be referred to as *objects* (or rich picture objects). In the current version of FSSE, a frame can have any number of tags, but in the next version each frame (i.e., each rich picture object) will have a single class. This implies that FSSE10 will only support single inheritance, which is a sensible solution to keep the tool simple. Moreover, in our analysis of past rich pictures we discovered that multiple inheritance is virtually never considered by students' during OOA.

Another change will be that each class in the new palette will contain typical instances, called *prototypes*. This idea originated from observing a particular pattern of use of FSSE during the test. A user would create some frames, give them names and tags, and cluster them in an empty area of the rich picture (an example of *spatial reasoning* within FSSE). Later the user will proceed to create new frames by cloning the ones in the cluster. The cluster itself can be considered as an extension to the FSSE palette. In FSSE10 we will therefore allow the user to drag a rich picture object (e.g., a drawing of a dog) from her rich picture into a class of her palette (the class "Dog"). The dragged object will then be referred to as a *prototype* of that class, i.e., a typical representative of the class. When a new object of the class is created (in this case a new dog) the prototype (i.e., the drawing of the dog) will be cloned, to provide an initial look for the newly created object. Proceeding in this way, the palette will contain more and more classes, each with its own prototypical objects, that the user stored during her exploration of the system concepts. A side-effect of supporting prototypes is that the palette becomes more *persistent* and easier to interpret even separated by the RP that generated it. This, in turn, opens the possibility of sharing a palette among many rich pictures, which is impossible in the current version.

### *B.Processes: arrows, events and conflicts*

Processes, a very relevant aspect of RP, are not directly supported in FSSE. In FSSE10 we plan to use *events* to represent processes. We already decided to provide labeled arrows, since they were explicitly required by our students in the test, so events will be implemented as a arrows between rich picture objects. Finally, conflicts will be considered as a special kind of events.

We are considering the possibility to implement events as *hyperedges*. Hyperedges are related to hypergraphs, a generalization of graphs [25]. A hypergraph can be defined as a set of vertices, and a set of hyperedges between the vertices; hyperedges are usually undirected, and represent relationship among 1 or more vertices. As an example, consider a FSSE10 user who wants to define an event "serve cake", involving 3 rich picture objects: a cake, a knife and a person. The user could select the objects and connect them via a single hyperedge labeled "serve cake". Each object attached to the hyperedge will have a *role*, specified by a *role name*; in the example the roles could be: "item to cut" for the cake, "cut with" for the knife, and "who" for the person. Roles of an event should be typed: e.g., the "item to cut" needs to be an object of the same class of the cake. An *event type* can later be created from the "serve cake" event, and attached to the palette. The event type will keep information about the role names and their required types, providing a mechanism to constraint and validate events. In the cake example, to serve a cake you need to link the role "who" to an object of class person, and FSSE10 should issue a warning if the role is attached to a dog.

Finally, in FSSE10 it would be easy to consider a conflicts as just another kind of events, i.e., labeled hyperedges among the parts of the rich picture that experience the conflict. However, we have noticed that conflicts tend to be neglected by our students, even if they are often necessary to make good RP. Therefore, we believe that our tool should provide an affordance for conflicts, for example in the form of a button for the specific creation of conflicts.

### *C.New file format*

A FSSE10 project will be a collection of rich pictures, together with a single, common palette (as depicted in Figure 6), and for this we need to define a new file format for FSSE10. The new format also reflects the special role and importance of the palette: it contains all ontological and behavioral information about the set of RP in a project. The palette also provides examples of typical objects of a domain (i.e., complete objects that serve as prototypes for the various classes), and data in natural language about conflicts and reflections around the rich pictures, in the form of concerns. We propose to consider the new palette as the initial core of a Domain Specific Language, in the sense expressed by Fowler [26]:

*"If people want to think about **[a system's]** behavior with events, states, and transitions—then we want that vocabulary to be present in the software code too. This is essentially the **Domain Driven Design principle of Ubiquitous Language**--that is we construct a shared language between the domain people **[...]** and programmers."*

This shared language in our case is a *visual shared language*, and the programmers should at least be able to use FSSE10 to agree among themselves, and whenever possible, with domain specialists and users too.

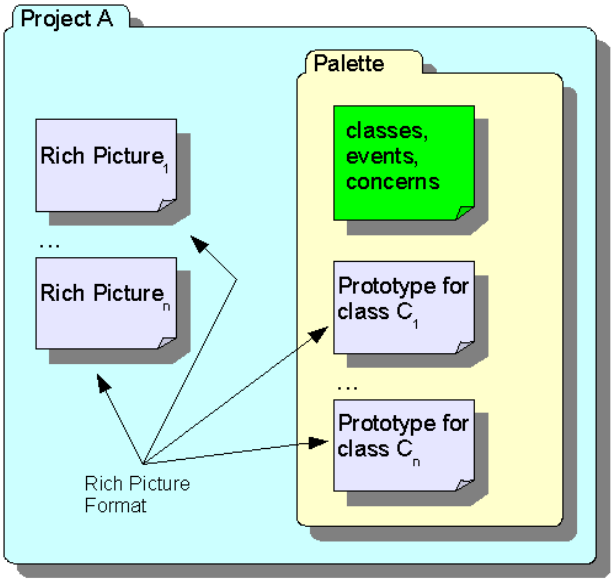



Figure 6: The new FSSE10 file format. A FSSE10 project is saved as a folder (labeled "Project A" in the figure, and colored cyan). Inside the project folder there is a sub-folder (yellow, labelled "Palette") which contains definition of classes, events and concerns. Some classes might have prototypes (i.e., examples of one or more common instances of that class), and those are also stored inside the palette folder, Moreover, in the project folder there is an XML file describing each individual rich picture. This storage format reflects the fact that all rich pictures in the same project share a common palette.

*D.Intelligence, flexibility and cooperation*

In FSSE we implemented a few algorithms to analyze the way the user nests her frame, and infer aggregation relationships among tags, as well as cardinality and optionality. In the next version we would like to provide mechanisms for discovery of contextual information: the *context* of a frame can be defined as the types its the surrounding frames. Relationships could be discovered using heuristics based on this notion of context.

We are also considering to improve the flexibility of our tool, by providing FSSE10 with a plug-in mechanism to enable users to define their own mapping from rich pictures to external formats, and perhaps to code.

From a social point of view, FSSE should be re-conceptualized in order to allow groups to actively interact with the program in their group rooms, and as it was suggested by one of our testers, also through the Internet from remote locations. It could be interesting to explore the effect of both synchronous and asynchronous virtual interaction.

*E.Mock-up of FSSE10*

To develop the new version of our RP authoring tool we are proceeding in an agile way, defining stories and selecting the most relevant ones to be the basis of the design and implementation incrementally more complex prototypes.

Since we advocate the use of RP in the analysis phase of software development, we sketched our stories to be visual and similar to rich pictures. Figure 7 shows the new look of the FSSE10 GUI, some of the steps in the creation of two rich pictures, about the same domain, and the incremental definition of a palette. The images in Figure 7 show, from top-right to bottom-left:

- the creation of visual representation for 3 objects: a house, a man and a car. The man is inside (a part of) the house. When the user assigns types to the 3 objects, the classes H (for the house), M (for the man) and C (for the car) are automatically added to the palette. The palette also detects that objects of class M can be inside objects of class H, and shows a 1-to-1 relationship between the 2 classes.
- the user creates an event called "sleep" that relates a man and his house. The role of the man is labeled "who" and the role of the house is "place".
- After creating the event "sleep", the user can declare an event type from the specific event. The "sleep" event type is added to the palette, at the bottom, and keeps information about the roles and their types: objects linked to the role label "who" should be of class M and objects with role "place" should be of class H. New events "sleep" can be created clicking on the event type in the palette.
- the user can set the object "house" as prototype of class H, by dragging it to the class H in the palette.
- now the user can save and close the current rich picture and start working on a fresh one, still keeping the same palette of classes and events. Populating the new rich picture should be quicker thanks to the knowledge in the palette. The user creates 2 new objects from class H, "house" and "myHouse". The "myHouse" object is a clone of "house" with some details altered. Class H uses its prototype to initialize each new instances.
- the user can declare that "me" sleeps in "myHouse", by creating a new event from event type "sleep", and linking the roles "who" to "me" and "place" to "myHouse". Finally a concern is created, shaped like a text bubble, in the top-right of the last image.

## VII. Conclusion

This paper describes the features and development of Free Sketch SE, a software tool to support rich pictures authoring for object-oriented analysis. To validate and complete the initial prototype of the tool, we ran a usability test. Although limited to a small group, the test provided meaningful feedback that is directing the next development iteration.

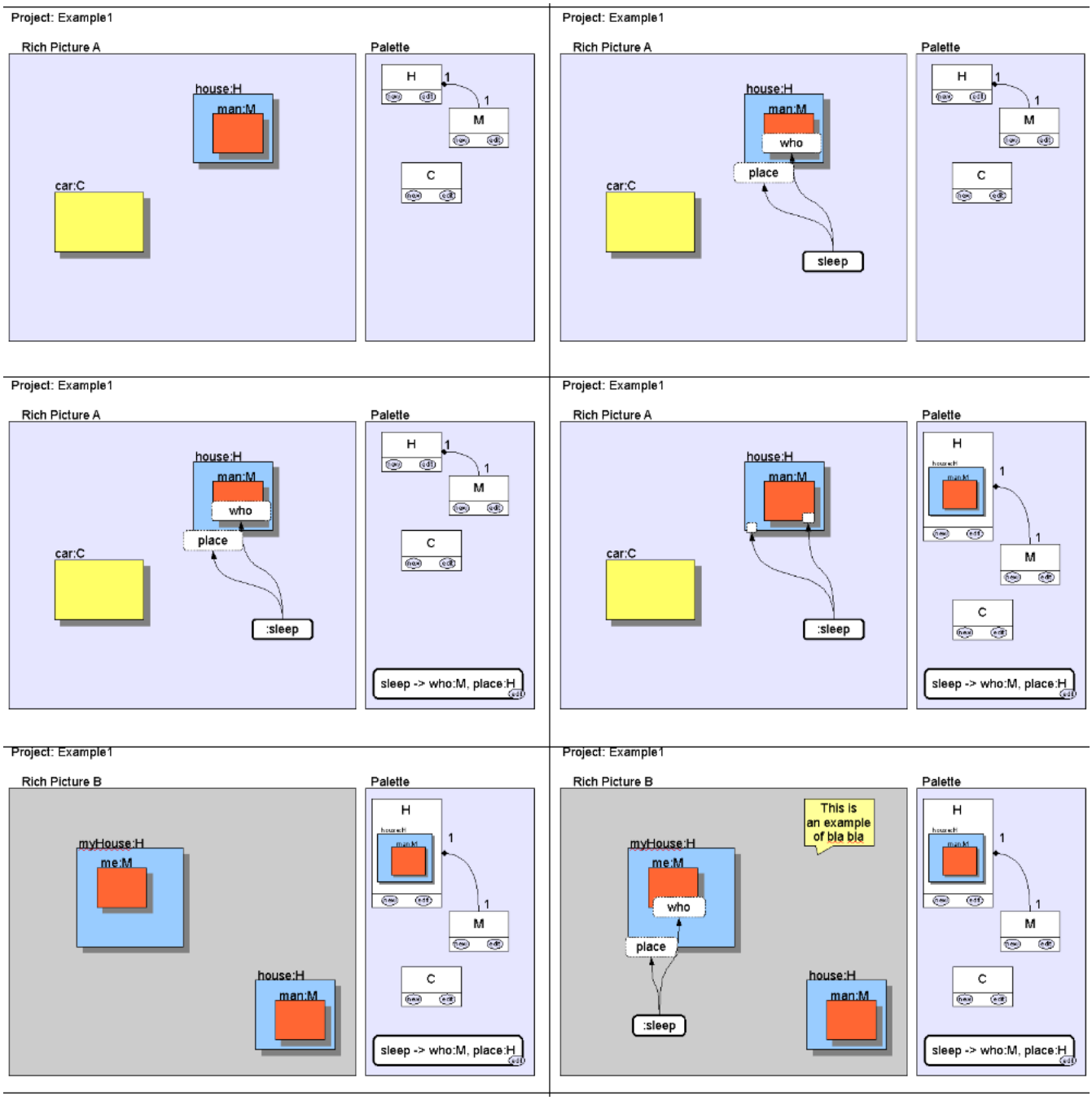



Figure 7: The new GUI of FSSE10. The images show (top-right to bottom-left) the progression of steps needed to create two rich pictures about the same problem. The palette is defined incrementally during the creation of the first rich picture; classes and events are specified and will be permanently stored in the palette. The second rich picture can then be built, leveraging on the elements already in the palette. Notice how all typical elements of a rich picture are now supported in FSSE: classes, events and constraints (bottom-left step).

After the test we reflected upon patterns of use and analyzed rich pictures in projects from various past semesters. From these we obtained a better understanding of how students create their rich pictures and what role they see for rich pictures in their project reports. The user-centered approach we followed proved of great help in better defining our tool's features: for example, the feedback received suggested us how to include support for events.

Moreover, we discovered something important about programmers and their values. They like to use authoring software tools at different phases of their project and they are happy to experiment with new ones. Furthermore, we realized that a software-supported activity makes immediately more sense to them and they are more willing to engage in it. They definitely dislike hand-drawing and try to avoid it. On a more general level, designerly activities, which are by nature *open*, are generally considered confusing and frustrating. An important lesson to keep in mind while developing designerly tools for programmers.

On the long run, we plan to improve Free Sketch, test it further, and deploy it as the main tool for a bachelor-level object-oriented analysis and design course.

## Acknowledgment

We thank the participants to our study and the anonymous referees who provided valuable suggestions to improve our paper.

## References

[1] A. Monk and S. Howard, “Methods & tools: the rich picture: a tool for reasoning about work context”. In Interactions, vol. 5, n. 2, pp 21-30, March 1998.

[2] K. Kotiadis and S. Robinson, “Conceptual modelling: knowledge acquisition and model abstraction.” In proceedings of the 40th Conference on Winter Simulation, Miami, Florida, pp. 951-958, December 07 - 10, 2008.

[3] L. Mathiassen, A. Munk-Madsen, P. A. Nielsen, and J. Stage, “Object-Oriented Analysis & Design”. Marko Publishing, ISBN: 87-7751-150-6, 1st edition, 2000.

[4] A. J. Cañas, R. Carff, G. Hill, M. Carvalho, M. Arguedas, T. C. Eskridge, J. Lott, and R. Carvajal, “Concept Maps: Integrating Knowledge and Information Visualization Export”. In Knowledge and Information Visualization journal, pp. 205-219, 2005.

[5] B. A. Nardi and J. A. Johnson, “User Preferences for Task Specific vs. Generic Application Software”. In Conference on Human Factors in Computing Systems CHI 1994, Boston, Massachusetts, USA, 1994.

[6] H. C. Mayr and C. Kop, “Conceptual Predesign - Bridging the Gap between Requirements and Conceptual Design”. In proceedings of the 3rd international Conference on Requirements Engineering: Putting Requirements Engineering To Practice, ICRE, IEEE Computer Society, Washington DC, April 06 - 10, 1998.

[7] E. Nuutila and S. Torma, “Text Graphs: Accurate Concept Mapping with Well-Defined Meaning”. In proceedings of the First International Conference on Concept Mapping, CMC 2004, Sept. 14-17, 2004.

[8] A. Valente, “Visual Middle-Out Modeling of Problem Spaces”. In International Conference on Information, Process, and Knowledge Management, pp. 43-48, February 01-07, 2009.

[9] A. Valente and E. Marchetti, "Please Don't Make Me Draw!: Lesson learned during the development of a software to support early analysis of object-oriented systems". In proceedings of the Second International Conference on Information, Process, and Knowledge Management (eKnow 2010) Saint Maarten, Netherlands, Antilles, pp 94-99, 2010.

[10] "VP-UML User's Guide" http://www.visual-paradigm.com/

[11] support/documents/vpumluserguide/12/13/5963_aboutvisualp.html Last visited 19 January 2011.

[12] Microsoft Visio. http://www.microsoft.com/office/visio/ Last visited 19 January 2011.

[13] DIA tutorial. http://live.gnome.org/Dia/Documentation Last visited 19 January 2011.

[14] H. Hsieh and F. Shipman, “Manipulating Structured Information in a Visual Workspace”. In proceedings of ACM Conference on User Interface Software and Technology, pp. 217-226, 2002.

[15] L. Cardelli and A. D. Gordon, “Mobile ambients”. In Theoretical Computer Science, vol. 240, Issue 1, pp. 177-213, June 6, 2000.

[16] J. Löwgren and E. Stolterman, “Thoughtful Interaction Design. A design perspective on information technology”. MIT Press, USA, 2005.

[17] S. Yliriksu and J. Buur, “Designing with video”. Springer, 2007.

[18] D. Schön, “The reflective practitioner. How professionals think in action”. Ashgate, London, UK, 1991.

[19] J. Preece, Y. Rogers, and E. Sharp, “Interaction Design. Beyond Human Computer Interaction”. John Wiley and Sons, USA, 2002.

[20] E. Stolterman, J. MacAtee, D. Royer, and S. Thandapani, “Designerly Tools”. In Undisciplined! proceedings of the Design Research Society Conference 2008, Sheffield, UK, pp. 116/1-14, July 2008.

[21] L. Suchman, J. Blomberg, J. E. Orr, and R. Trigg, “Reconstructing Technology as Social Practice”. In American Behavioral Scientist, vol. 43 n. 3, Sage Publications, pp. 392-408, November/December 1999.

[22] L. Suchman, “Affiliative Objects”. In Organizations 2005, vol. 12, n. 3, Sage Publications, pp. 379-399, 2005.

[23] H. W. J. Rittel and M. M. Webber, “Dilemmas in a general theory of planning”. In Policy Sciences, n. 4, 1973, first edition American Association for the Advancement of Science, Boston USA, pp. 155-169, December 1973.

[24] T. Ingold and E. Hallam, “Creativity and Cultural Improvisation”. Berg Publishers, 2008.

[25] A. Agarawala and R. Balakrishnan, “Keepin' it real: pushing the desktop metaphor with physics, piles and the pen”. In proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '06, ACM, New York, NY, Montréal, Québec, Canada, pp. 1283-1292, April 22 - 27, 2006. DOI http://doi.acm.org/10.1145/1124772.1124965

[26] F. Drewes, B. Hoffmann, and D. Plump, “Hierarchical graph transformation”. In Journal of Computer and System Sciences, vol. 64, n. 2 pp. 249-283, March 2002. DOI http://dx.doi.org/10.1006/jcss.2001.1790

[27] M. Fowler and R. Parsons, “Domain Specific Languages”. Addison-Wesley Professional, 2010. ISBN: 0321712943.

# Aggregating Geoprocessing Services using the OAI-ORE Data Model

Carlos Abargues, Carlos Granell, Laura Díaz, Joaquín Huerta
*Institute of New Imaging Tecnologies (INIT)*
*Universitat Jaume I (UJI)*
*Castellón, Spain*
{*abargues, carlos.granell, laura.diaz, huerta*}*@uji.es*

***Abstract*—Rapid discovery and access of geospatial resources is critical for many application domains that require agile data integration. In this context, cross-domain geospatial applications need immediate access to geospatial resources of interest in order to rapidly integrate them in scalable, functional Web applications. In this paper we explore new perspectives to build pragmatic geospatial Web applications, drawing on the ideas of recent initiatives like Linked Open Data and Open Archives Initiative. By using and extending standards and principles from these initiatives we are able to model single and composite geoprocessing services as a collection of heterogeneous Web resources. Such collections are built by the principle aggregation by linking that enables to connect and link multiple geospatial data and services across different application domains.**

***Keywords*-models for geocomputation; geoprocessing services; collections; service integration; OAI-ORE; Linked Open Data**

## I. INTRODUCTION

Geospatial data sets are increasingly becoming available in open repositories. Not only as official, validated data sets collected by authorities and experts that make them available through catalogues in Spatial Data Infrastructures (SDI) nodes, but also as on-line resources dispersed everywhere produced by thousands of individuals. Nonexpert users are taking the role of productors of geospatial data through the massive use of social networks (Flickr, Twitter, etc.) and location-based devices (mobile phones, digital cameras, etc.), which leads to huge amounts of rich georeferenced user-generated content in a great variety of sizes and formats [2]. As some authors pointed out [3], SDI nodes cannot ignore the millions of users providing up-to-date data anywhere at any time, becoming a challenging task for the next generation of geospatial applications to conciliate the up-bottom approach in traditional SDI creation with the bottom-up approach powered by users [4].

In the SDI context geoprocessing services are a powerful means for creating web-based applications, integrating geospatial data from multiple sources [5] [6] [7]. In addition, Brauner et al. [8] have recently reported a set of research projects focused on designing and implementing geoprocessing services, in order to draw a research agenda for future developments in the realm of distributed geoprocessing computing. According to these studies [3] [8], some issues that should be addressed in the near future are:

- The need of integrating information from multiple and heterogeneous sources to benefit from the rich, valuable, up-to-date user-generated content.
- The lack of flexible mechanisms to create on-demand, scalable service-oriented geospatial solutions that contains relevant geospatial resources such as geoprocessing services, geospatial data and user-generated content.

Our long-term research goal pursues new perspectives to provide pragmatic, scalable approaches to creation of geoprocessing workflows that take into account user-generated content out of SDI context. Here we focus on the facet of description (leaving other facets like execution aside for now) of geoprocessing services in such a way that let users create collections of related geoprocessing services.

The proposed approach draws on the principles and best practices exposed by Linked Open Data (LOD) [9] [10]. This initiative stems from the very principles of Web architecture and pursues the goal of "enabling people to share structured data on the Web as easily as they can share currently Web resources" [9]. The idea consists of publishing data in a structured manner and creating typed links between data resources from heterogeneous sources, assuming these two tenets: (i) RDF (Resource Description Framework) data model is used as common model to publish structured data; and (ii) extensive use of typed links to connect data from different data sources. Essentially, it is assumed that the more interlinked data, the more aspects of meaning (richness) might be represented.

Although accessing to data itself may be viewed as a prior goal, users are already retrieving georeferenced data either through SDI data download services or from open repositories such as Flickr, Open Street Maps and Twitter. The availability and stability of latter data sets are fundamental and should be reinforced, though, the real impediment is to provide better connections (interlinking) among the vast amount of user-generated data, SDI content, and geoprocessing services so that experts and nonexperts alike may access to structured data and services for their cross-domain geospatial applications. Furthermore, our aim here is to apply the OAI-ORE (Open Archive Initiative -

Object Reuse and Exchange) protocol (see Section II-C), which has been proven to be successful in digital library projects, extending it with new types of relationships to the geospatial landscape in order to support not only structured data, as LOD states, but also interlinking other types of Web resources like geoprocessing services.

This paper is an extended version of a conference paper referenced in [1]. The rest of the paper is structured as follows. In Section II we introduce the basic concepts used throughout this paper focusing specifically on the OAI-ORE's abstract data model. The proposed approach to model geoprocessing services and data as collections of interlinked heterogeneous resources is described in Section III. Section IV compares our work with related projects. Finally, Section V concludes by summarizing the key features of our approach and discussing ongoing work.

## II. Background

This section presents relevant concepts and definitions for the remaining sections. In particular, the following subsections will shortly introduce the SDI and related geospatial services, the LOD project, and the OAI-ORE specification.

### A. SDI and geoprocessing services

Spatial Data Infrastructures (SDI) describe the notion of service-oriented management, accessing, and processing of geospatial data. The implementation of SDI has traditionally followed a service oriented architecture paradigm where web services technology plays a enabling role in data integration and promotion of interoperability among heterogeneous distributed information sources [11].

Geospatial web services allow users to access, manage, and process geospatial data in a distributed manner [12]. The demand for interoperability has boosted the development of standards and tools to facilitate data transformation and integration, mostly in terms of standard interfaces specified by Open Geospatial Consortium (OGC[1]) and Technical Committee 211 (TC211[2]) of International Organization for Standardization (ISO). The Web Map Service (WMS), the Web Feature Service (WFS) and the Web Coverage Service (WCS) are some prominent examples of OGC interfaces for geospatial services. All come in different versions, where WMS 1.3.0 [13], WFS 1.1.0 [14], and WCS 1.1.2 [15] are the most recent. The central building-blocks for data, as well as service discovery, are provided by the Catalogue Services for the Web (CSW) [16] and so called geoportals [17]. The CSW provides one access point to users that search for geospatial data.

Geoprocessing services essentially transform geospatial data to produce new data or meaningful information [18]. A substantial leap ahead in the domain of geoprocessing services was the OGC Web Processing Service (WPS) specification [19]. This specification was designed to encapsulate generic operations and algorithms over the Internet. The basic operational unit of the OGC WPS is the notion of process, that is, a geospatial operation with inputs and outputs of a defined type. This means that a given WPS instance (a concrete WPS service running) may offer one or various operations (or processes) as normal web services do. The common communication pattern between a client and a WPS instance encompasses three types of requests. A request can be sent to the WPS instance via HTTP GET with parameters provided as Key-Value Pairs (KVP) or via HTTP POST, with parameters supplied in a XML document. These three types of requests are:

- *GetCapabilities*. First, a WPS instance receives a KVP *getCapabilities* request (which is common for all OGC geospatial services) and simply responds with an XML document, containing metadata such as server provider, contact information, general description, and a list of contained geoprocessing operations (processes) offered by the queried WPS instance.
- *DescribeProcess*. A WPS-client selects a process identifier from the *getCapabilities* response and performs a *describeProcess* request, either as KVP or as XML document. The WPS instance responds with an XML document containing needed information for the solicited process, such as input and output parameter names and types, so that the WPS-client may later build the execute request.
- *Execute*. The WPS-client eventually requests the execution of a geospatial operation, with all required input data by invoking the execute method as an XML document request. The WPS instance then runs the operation and returns the results informing also of its status.

As geospatial web solutions continue to grow and increase in complexity, many standards organizations, industry bodies, and the geospatial research community have paid attention to the effective composition and orchestration of geospatial web services [20]. Rather than describing a new service interface for geoprocessing services, our aim here is to propose a new way to improve service compositions in terms of collections or aggregations of geoprocessing services in line with the principles of the Linked Data community.

### B. Linked Open Data

Linked Open Data (LOD) represents a style of information publishing on the Web. This style relies on traditional web technologies and the usage of light-weight techniques for data model representation. The former resides on the use of Uniform Resource Identifiers (URI) as reference points. A URI is used to uniquely identify a resource, i.e., a piece of data, and also for actual access to the resource

[1] http://www.opengeospatial.org/
[2] http://www.isotc211.org/

representation. This implies that HTTP URI should be deferenciable URI, that is, user can look up these URIs to retrieve resource representations. Content negotiation comes here to allow clients to specify an acceptable representation of a data set [21]. While connecting to a data source, the client may specify the desired representation. This may be, for instance, plain RDF, or an HTML representation with increased readability for the human user.

The former refers to the Resource Description Framework (RDF) as basic structure for any form of description. RDF provides means to describe any kind of resource in form of triples (subject-predicate-object). In this way, data published according to LOD principles is exposed in RDF format and interlinked by exploting the intrinsic capabilities of the RDF model to link resources. As we will see in the next section, OAI-ORE and LOD share some characteristics what make OAI-ORE a suitable candidate to model collections of geoprocessing services with the benefits of the LOD project.

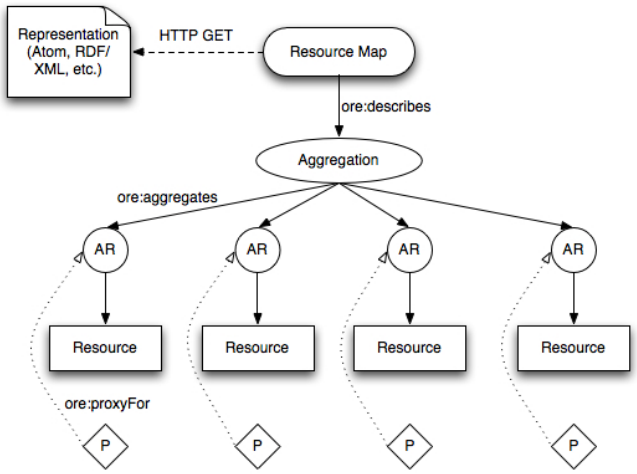



Figure 1. Simplified OAI-ORE's abstract data model

### C. OAI-ORE's abstract data model

The Open Archive Initiative - Object Reuse and Exchange (OAI-ORE) protocol [22] defines an abstract data model [23] for describing, reusing, and exchanging collections of Web resources. The aim of this protocol is to expose rich content (text, images, data, video, etc.) in aggregations to be then fed by applications in the realm of digital library domain. Obviously, OAI-ORE is closely related with the OAI - Protocol for Metadata Harvesting (OAI-PMH) [24], since for instance source content (e.g., e-prints records) described in OAI-ORE can be harvested automatically in order to replicate it in others remote repositories [25].

Conceptually, the OAI-ORE's abstract data model builds strongly on the principles defined on LOD. First, the notion of "addressable resources" indicates that resources of any type (file, image, text document, metadata, process, etc.) should be identified using HTTP URI. Secondly, exploiting the simple mechanism of "typed links" to connect resources enables the discovery, browsing, and access to more related and connected data.

In principle, the use of the capabilities of OAI-ORE and LOD to build scalable, distributed Web applications that integrate heterogeneous remote sources becomes evident [26]; LOD takes the principles of the current Web architecture, the most, by far, scalable and distributed information system. As the OAI-ORE's abstract data model relies on such principles, we find a rational argument to use it to link collections of Web resources, considering a geoprocessing service as a particular type of Web resource.

Before discussing how the OAI-ORE's abstract data model can be used for describing and linking geoprocessing services (see Section III), we introduce here its key entities [23]. The simplified diagram in Figure 1 shows how these entities are related each other. The entity Aggregation plays a central role as it represents a collection of addressable resources that in turn are called Aggregated Resources (AR). The ore:aggregates relation denotes here the "aggregation by linking" mechanism to connect resources related in some way. This implies that both Aggregations and AR entities are addressable resources in the sense that both use HTTP URI as referencing method. Here comes the process of dereferencing URI, another key aspect in LOD, which just means looking up a URI on the Web in order to get either the resource itself or its representation.

Aggregation and AR are abstract terms that must still refer to concrete resources, which can be of any type such as a document, image, process, service, and even a chain of geoprocessing services as we will see in the following section. The OAI-ORE specification makes use of the Resource Map entity to provide a concrete representation for the whole aggregation, mostly derived from RDF. Some suggested formats in the specification are the Atom syndication format [27], RDF/XML[3], and RDFa[4] (a microformat for extending XHTML to support RDF). We use the RDF/XML serialization so that resulting collections can be readily added and connected to other LOD datasets since RDF triples are the common data model.

OAI-ORE defines a useful abstract entity called Proxy (P) by which it is possible to express the role an aggregated resource has explicitly in the context of an aggregation. For example, two resources may have a temporal relationship that connects to each other and this is only meaningful within the aggregation context in which they are defined. The use of relationships from and to Proxy elements instead of the Aggregated Resource elements they represent does not affect the original resource. In addtion, the use of Proxy elements

[3]http://www.w3.org/TR/rdf-syntax-grammar/
[4]http://www.w3.org/TR/rdfa-syntax/

does not make private information explicit (e.g., temporal and semantic relationships) to external aggregations because either this knowledge is not required or is meaningless to them. The Proxy entity thus enables encapsulation, a major driving force to support reusability [28].

Besides the Proxy entity, OAI-ORE protocol permits the use of relationships to link directly to resources and aggregations. These relationships can be either internal, between the resources defined within the aggregation, or external, linking to external resources such as georeferenced user-generated content. In both cases, specific relationships such as ore:aggregates and ore:describes are defined. An example of external relationship is ore:similarTo. However, following the "typed link" principle of LOD, those resources defined in a given OAI-ORE aggregation may, indeed should, link to and be linked from other external resources based on relationships characterized semantically by other vocabularies.

Aside from linking related resources, the "aggregation by linking" mechanism makes it ease to create complex hierarchical aggregations from simpler ones. This implies that collections of resources can be scaled and reused easily, since incorporating or eliminating a resource from a given collection simply means to refer or not to its HTTP URI. Next section focuses on how geoprocessing services can be seen as resources and described using the OAI-ORE's abstract data model.

## III. Approach

This section first proposes the conceptual architecture that supports our approach and lists some assumptions that drive our research at the present stage. Then a set of new relationships to model geoprocessing services inside an OAI-ORE colection is presented as an extension on this protocol. Finally we describe how to use the OAI-ORE and the new extension to model geoprocessing service and their composition.

### A. Architecture

SDI-based applications are built upon a multilayer architecture as depicted in the right side of Figure 2 (blue boxes). Application layer may contain thin client tools such as geoportals, mashups and rich internet applications (e.g., Flex, JavaFX), and also thick desktop-based clients. The middleware layer comprises multiple distributed services, which allow client applications to discover, access, and process geospatial data and metadata from remote repositories (Data layer). SDI applications work in this way because SDI nodes are normally architectured in such a way[5].

To provide better connections with other type of data out of SDI content, like user-generated content and LOD datasets (blue clouds and gray circles in Figure 2), we have added a Transformation process to convert heterogeneous

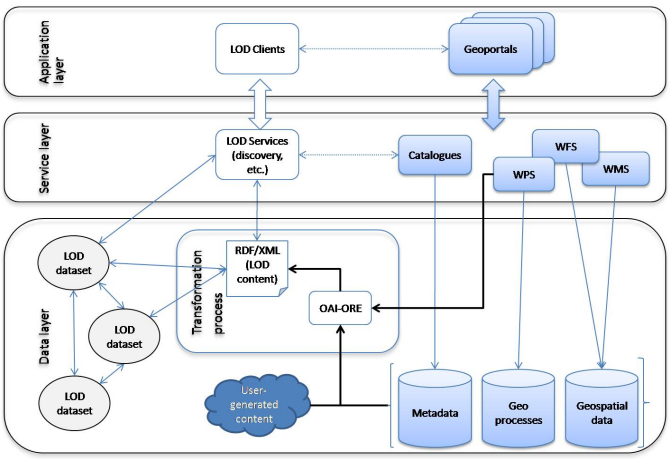



Figure 2. Proposed architecture

source resources into RDF/XML format. The source resources may be geospatial data sets in SDI repositories, georeferenced documents, pictures or tweets (Twitter messages), as well as geoprocessing services to compute calculations over such datasets. The black arrows in Figure 2 shows the two-step transformation process: first resource collections are generated based on the OAI-ORE data model to be then serialized as RDF/XML documents. Besides resulting collections may be readily connected to resources in other RDF-based LOD datasets[6] across several communities and domains [9].

The proposed architecture poses some requirements for interlinking geoprocessing resources: addressable resources (data and services), connecting (linking) resources, and the ability to see and link resource descriptions (metadata) for each resource. In the following we enumerate some assumptions taken in this paper for describing collections of interlinked Web resources:

- Geospatial datasets and services are addresable resources, i.e., have referenciable HTTP URI [9] [10] [26].
- A geoprocessing service is considered a type of Web resource, which let us rely on the OAI-ORE protocol to model interlinking geoprocessing services.
- Since a geoprocesing service is a resource, a chain of geoprocessing services may be then comparable to a collection of heterogeneous Web resources. A collection thereby is a resource with own metadata.

### B. Extending the OAI-ORE abstract data model

A geoprocessing service can be characterized by its signature: a function or capability, and a set of input and output parameters. Taking this simple approach, the mapping to the OAI-ORE's abstract data model is driven by the following

[5]http://inspire.jrc.ec.europa.eu/reports/ImplementingRules/network/D3_5_INSPIRE_NS_Architecture_v3-0.pdf

[6]http://richard.cyganiak.de/2007/10/lod/

rules: (i) a geoprocessing service is an Aggregation entity; (ii) its components are Aggregated Resources, i.e., the capability of the resource, and the collections of input and output parameters; (iii) a Resource Map entity then describes the Aggregation using a serialization format.

These assumptions describe a method for mapping geoprocessing service elements with the OAI-ORE data model, however, the built-in relationships defined by this specification are not enough to express the roles and relationships for modelling aggregations of geoprocessing services. In order to appropiately define these relationships, we have extended the OAI-ORE abstract data model with a set of new relationships (with prefix *ores*) that can be applied to the already existing elements.

Table I briefly shows these new relationships altogheter with their name, URI, the inverse relationship in the case this exists and their domain and range properties. Not only the definition of these relationships but also the elements that make use of them is an important aspect to consider. The relationships among the different elements that describe a geoprocessing service can only be meaningful in the context of this description and they may not be useful or representative outside the geoprocessing service description for an aggregated resource. In order to encapsulate this information most of the new relationships defined are applied to the Proxy element. The Proxy actually becomes a key element for describing the geoprocessing service representing and containing the internal relationships that models it.

Geoprocessing services may receive an input to generate an output or result. This behaviour is modeled in our approach through the use of the relationships ores:aggregatesInput, ores:inputAggregatedBy, ores:aggregatesOutput and ores:outputAggregatedBy. The first two relationships allow the specification of the input for a given service being each relationship the inverse for the other. The last two ones do the proper in the task of specifying the output for the service representing again an inverse relationship among them. Since a service can have zero or more inputs (and outputs), the cardinality of these relationships can be specified as zero to any.

Similar behaviour occurs when defining the inner processes that compounds the geoprocessing service. This components can also specify their inputs and outputs using the declared relationships ores:inputFor, ores:hasInput, ores:outputFor, ores:hasOutput. A given process may have zero or more inputs and zero or more outputs that can be specified by using any number of relationships to link both, the process and the inputs or outputs represented by their corresponding Proxy element.

The different processes that compound the geoprocessing service have to be described in a certain partial order. To model this feature, the proposed extension of the OAI-ORE data model defines the relationships ores:next and ores:previous. These two new relationships can be applied only to those Proxy elements that represent processes as aggregated resources and indicate the next or the previous processes that should be followed during the service execution. Thanks to these relationships it is possible not only to specify an order for the inner processes but also navigate among them and check for instance the requirements (i.e., output of a process that serves as input for another) in order to start the execution of a process.

OAI-ORE defines a method for reusing defined aggrega-

Table I
OAI-ORE'S MODEL EXTENSION FOR SERVICE DESCRIPTION

| | |
|---|---|
| *Name* | ores:**aggregatesInput** |
| *URI* | http://www.geoinfo.uji.es/ores/terms/aggregatesInput |
| *Inverse* Of | ores:inputAggregatedBy |
| *Domain* | ore:Aggregation |
| *Range* | ore:Proxy |
| *Name* | ores:**inputAggregatedBy** |
| *URI* | http://www.geoinfo.uji.es/ores/terms/inputAggregatedBy |
| *Inverse* Of | ores:aggregatesInput |
| *Domain* | ore:Proxy |
| *Range* | ore:Aggregation |
| *Name* | ores:**aggregatesOutput** |
| *URI* | http://www.geoinfo.uji.es/ores/terms/aggregatesOutput |
| *Inverse* Of | ores:outputAggregatedBy |
| *Domain* | ore:Aggregation |
| *Range* | ore:Proxy |
| *Name* | ores:**outputAggregatedBy** |
| *URI* | http://www.geoinfo.uji.es/ores/terms/outputAggregatedBy |
| *Inverse* Of | ores:aggregatesOutput |
| *Domain* | ore:Proxy |
| *Range* | ore:Aggregation |
| *Name* | ores:**inputFor** |
| *URI* | http://www.geoinfo.uji.es/ores/terms/inputFor |
| *Inverse* Of | ores:hasInput |
| *Domain* | ore:Proxy |
| *Range* | ore:Proxy |
| *Name* | ores:**hasInput** |
| *URI* | http://www.geoinfo.uji.es/ores/terms/hasInput |
| *Inverse* Of | ores:InputFor |
| *Domain* | ore:Proxy |
| *Range* | ore:Proxy |
| *Name* | ores:**outputFor** |
| *URI* | http://www.geoinfo.uji.es/ores/terms/outputFor |
| *Inverse* Of | ores:hasOutput |
| *Domain* | ore:Proxy |
| *Range* | ore:Proxy |
| *Name* | ores:**hasOutput** |
| *URI* | http://www.geoinfo.uji.es/ores/terms/hasOutput |
| *Inverse* Of | ores:outputFor |
| *Domain* | ore:Proxy |
| *Range* | ore:Proxy |
| *Name* | ores:**next** |
| *URI* | http://www.geoinfo.uji.es/ores/terms/next |
| *Inverse* Of | ores:previous |
| *Domain* | ore:Proxy |
| *Range* | ore:Proxy |
| *Name* | ores:**previous** |
| *URI* | http://www.geoinfo.uji.es/ores/terms/previous |
| *Inverse* Of | ores:next |
| *Domain* | ore:Proxy |
| *Range* | ore:Proxy |

tions in others as aggregated resources of the latter indicating the resource map or serialization of the former by the relationship ore:isDescribedBy that links both. The same may happen when describing a geoprocessing service, however, in this case the aggregations can have two different roles: either describing a collection of aggregated resources or describing a executable geoprocessing service. In the first case all of the aggregated resources could be used as input for a given process. In the second case, through, it makes sense that only the aggregated resources that represent the output of its execution may be of interest instead of the entire service description (collection).

This behaviour represents an ambiguity concerning the role a nested aggregation plays in a service description. To resolve it the relationship ores:hasOutput can be used to specify in both cases those elements that can be reused by other geoprocessing service descriptions when the first aggregation is used as one of its aggregated resources. The use of ores:hasOutput makes it easy to reuse the output of a geoprocessing service or a collection of other resources and also allows to specify only a subset of the aggregated resources in a collection or intermediate results in a geoprocessing service that may be of interest although they are not the result for the service execution.

### C. Modeling geoprocessing services as Web resources

Considering the OAI-ORE specification and the previously explained extension based on it, let us consider now a concrete geoprocessing service like a transformation service that converts a source KML (Keyhole Markup Language) file [29] into a GML (Geography Markup Language) format [30]. This service is called Kml2Gml and takes one input parameter –a data resource–, and returns the corresponding GML content as an addressable resource so that can be retrieved by dereferencing its URI.

Figure 3 illustrates this simple scenario using a named graph to represent the mappings between the Kml2Gml service and the OAI-ORE's abstract data model including the relationships defined in the extension for describing geoprocessing services. From top to bottom, the Resource Map entity named ReM-1 describes, through the relationship ore:describes, the Aggregation entity A-1. The A-1 entity is composed, through the relationship ore:aggregates, of three Aggregated Resource entities, named AR-1, AR-2, and AR-3 respectively. These AR entities map to counterpart resources of the Gml2Kml service. In this particular case, AR-2 represents the function, AR-1 the input resource, and AR-3 the output resource. Their corresponding Proxy elements are also represented in the graph by the elements P-1, P-2 and P-3 respectevily and allow the definition of the required relationships for the aggregated resources that are meaningful only in the service description context. Table II lists the set of URL for the resulting addressable resources, both abstract (Aggregation, Aggregated Resource and Proxy

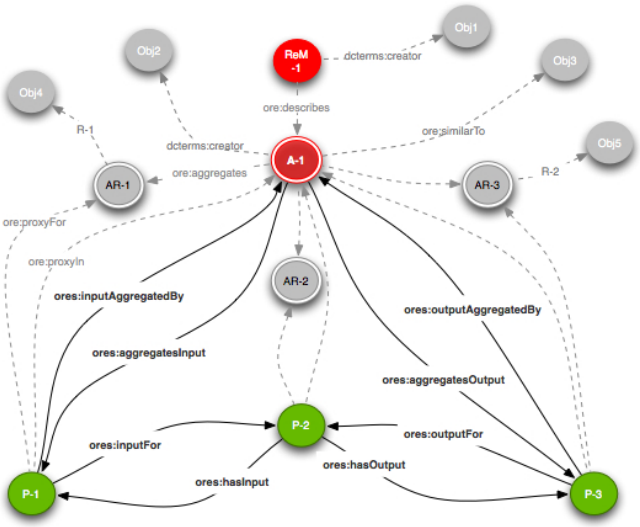



Figure 3. Mapping a geoprocessing service in OAI-ORE

Table II
LIST OF ADDRESSABLE RESOURCES

| ENTITIES | URI |
|---|---|
| ReM-1 | http://www.geoinfo.uji.es/resource/aggregation.rdf |
| A-1 | http://www.geoinfo.uji.es/resource/aggregation |
| AR-1 | http://www.geoinfo.uji.es/data/datasetKML.kml |
| AR-2 | http://www.geoinfo.uji.es/process/Kml2Gml |
| AR-3 | http://www.geoinfo.uji.es/data/datasetGML.gml |
| P-1 | http://www.geoinfo.uji.es/proxy/r?<br>what=http://www.geoinfo.uji.es/data/datasetKML.kml&<br>where=http://www.geoinfo.uji.es/resource/aggregation |
| P-2 | http://www.geoinfo.uji.es/proxy/r?<br>what=http://www.geoinfo.uji.es/process/Kml2Gml&<br>where=http://www.geoinfo.uji.es/resource/aggregation |
| P-3 | http://www.geoinfo.uji.es/proxy/r?<br>what=http://www.geoinfo.uji.es/data/datasetGML.gml&<br>where=http://www.geoinfo.uji.es/resource/aggregation |

entities) and concrete resources such as geospatial data files and services.

Figure 3 shows in grey tone all those relationships not defined by the presented extension. These relationships include those externally defined such as dcterms:creator, used by the Resource Map and Aggregation entities to point to an author (external) resource based on the Dublin Core vocabulary[7].

As described previously, AR-1 and AR-3 entities represent the needed input and output resources for the capability resource AR-2. To capture these relationships between resources in the OAI-ORE's abstract data model, it is required a couple of steps. First, we create a Proxy entity for each AR entity (P1, P2, and P3 respectively). So a Proxy entity is tied to the corresponding AR entity by two OAI-ORE built-in relationships, namely ore:proxyFor and ore:proxyIn. The former indicates that the Proxy entity is for a concrete

[7]http://dublincore.org/documents/dces/

Table III
LIST OF ADDRESSABLE RELATIONS

| RELATIONS | URI |
|---|---|
| ore:similarTo | http://www.openarchives.org/ore/terms/similarTo |
| ore:describes | http://www.openarchives.org/ore/terms/describes |
| ore:aggregates | http://www.openarchives.org/ore/terms/aggregates |
| ore:proxyFor | http://www.openarchives.org/ore/terms/proxyFor |
| ore:proxyIn | http://www.openarchives.org/ore/terms/proxyIn |
| dcterms:creator | http://purl.org/dc/terms/creator |
| ores: inputAggregatedBy | http://www.geoinfo.uji.es/ores/terms/ inputAggregatedBy |
| ores: aggregatesInput | http://www.geoinfo.uji.es/ores/terms/ aggregatesInput |
| ores: outputAggregatedBy | http://www.geoinfo.uji.es/ores/terms/ outputAggregatedBy |
| ores: aggregatesOutput | http://www.geoinfo.uji.es/ores/terms/ aggregatesOutput |
| ores:inputFor | http://www.geoinfo.uji.es/ores/terms/inputFor |
| ores:hasInput | http://www.geoinfo.uji.es/ores/terms/hasInput |
| ores:outputFor | http://www.geoinfo.uji.es/ores/terms/outputFor |
| ores:hasOutput | http://www.geoinfo.uji.es/ores/terms/hasOutput |

AR entity. The latter is informative and just keeps informed the upper Aggregation A-1 about the active Proxy entities. Second, semantic relationships among resources are possible by means of relationships belonging to vocabularies, either in the OAI-ORE domain itself or in others.

It is here where the extension plays its role modelling the geoprocessing service through the use of the requiered relationships. For example ores:hasInput and ores:inputFor indicate that AR-1 (input resource) serves as input to the AR-2 (capability resource). Something similar happens for the ores:hasOutput and ores:outputFor relationship but involving the AR-2 and AR-3 entities to indicate the output for the former. The relationships ores:inputAggregatedBy and its invers ores:aggregatesInput indicate that the entity AR-1 through its Proxy entity P-1 serve as input for the aggregation A-1 that models the entire service. In order to indicate the output of the service the relationships ores:outputAggregatedBy and ores:aggregatesOutput links both the Aggregation A-1 and the Aggregated Resource AR-3 through its Proxy P-3.

Although OAI-ORE data model and the proposed extension define the required relationships to model geoprocessing services as aggregations, other relationships and terms can be used to enrich a description of a collection. These terms and relationships can be created for a concrete purpose or reused from common vocabularies in other fields. The use of common vocabularies in the geoprocessing context is a tricky question. Some authors claim the need of a geoprocessing taxonomy [8], as a mechanims to provide a common semantic background on which discovery, access, and composition of geoprocessing services can be achieved. In other words, for effective interoperability it is a must to build and label similar resources in compatible ways. Otherwise, the task of choosing, extending, and merging vocabularies becomes a sensitive issue [31].

As a best practice, semantic terms should be reused from well-known vocabularies wherever possible, avoiding the definition of new terms if they already exist. As OAI-ORE protocol supports the use of existing vocabularies, we have followed this strategy by reusing relationships from well-known vocabularies like RDF schemas[8] (e.g., rdf:domain). In this way, additional metadata for the whole collection (e.g., use case, context, data provenance) may also be encoded as external resources using meaningful relationships. Table III lists the most relevant relationships used in the Km2Gml geoprocessing service scenario.

Another example of widely-used vocabulary with geographic connotations is the case of the vocabulary or ontology offered by Geonames[9]. This vocabulary is expressed in OWL [32] and offers a collection of over six million of place names and other relevant terms to express relationships. In addition, Geonames features are interlinked each other by means of typed links denoting hierarchical inclusion (e.g., continent, countries, administratives units, etc.) and proximity. The Geonames vocabulary seems to be a feasible choice to link geospatial resources (user-generated content, SDI content, OAI-ORE aggregations), thus offering a simple mechanism of georeferencing heterogeneous Web resources. Therefore, the combination of Proxy entities and meaningful relationships from (existing) vocabularies enables the definition of flexible and customized connections among disparate resources of a given aggregation.

Finally the code in Figure 4 shows a portion of a possible RDF/XML representation that describes a geoprocessing service as represented in Figure 3 following the OAI-ORE recommendations for serializing aggregations.

### D. Interlinking geoprocessing services

The previous section described how OAI-ORE's abstract data model is used and extended to describe a geoprocessing service. To illustrate our approach a simple geoprocessing services was used as example of aggregation of resources. Independently of its complexity any geoprocessing service can be modeled following the same pattern: a capacity resource, one or more input resources and one or more output resources. This section sketches how a chain of geoprocessing services is modelled as a collection of interlinked aggregated resources and how another collection can be reused as input for a capacity resource.

Figure 5 shows a chain of two services represented in OAI-ORE. In this case, the AR-2 and AR-4 entities represent geoprocessing services. Returning to our simple scenario, AR-2 refers to the Kml2Gml service while the AR-4 would be a topology function like intersection that operates over the results of AR-2 and a collection of geometries defined in AR-5. Again, the key aspect is the combination of Proxy entities and suitable relationships that express properties

[8]http://www.w3.org/TR/rdf-schema/
[9]http://www.geonames.org/ontology

```
<rdf:Description rdf:about="http://www.geoinfo.uji.es/resource/aggregation.rdf">
  <rdf:type rdf:resource="http://www.openarchives.org/ore/terms/ResourceMap"/>
  <ore:describes rdf:resource="http://www.geoinfo.uji.es/resource/aggregation"/>
</rdf:Description>
<rdf:Description rdf:about="http://www.geoinfo.uji.es/resource/aggregation">
  <rdf:type rdf:resource="http://www.openarchives.org/ore/terms/Aggregation"/>
  <ore:similarTo rdf:resource="http://www.geoinfo.uji.es/aggregations/aggr2"/>
  <dc:creator rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
    http://www.geoinfo.uji.es/personal/cabargues
  </dc:creator>
  <ore:isDescribedBy rdf:resource="http://www.geoinfo.uji.es/resource/
    aggregation.rdf"/>
  <ore:aggregates rdf:resource="http://www.geoinfo.uji.es/data/datasetKML.kml"/>
  <ore:aggregates rdf:resource="http://www.geoinfo.uji.es/process/Kml2Gml"/>
  <ore:aggregates rdf:resource="http://www.geoinfo.uji.es/data/datasetGML.gml"/>
  <ores:AggregatesInput rdf:resource="http://www.geoinfo.uji.es/proxy/r?
    what=http://www.geoinfo.uji.es/data/datasetKML.kml&
    where=http://www.geoinfo.uji.es/resource/aggregation"/>
  <ores:AggregatesOutput rdf:resource="http://www.geoinfo.uji.es/proxy/r?
    what=http://www.geoinfo.uji.es/data/datasetGML.gml&
    where=http://www.geoinfo.uji.es/resource/aggregation"/>
</rdf:Description>
<rdf:Description rdf:about="http://www.geoinfo.uji.es/data/datasetKML.kml">
  <ore:isAggregatedBy rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
    http://www.geoinfo.uji.es/resource/aggregation
  </ore:isAggregatedBy>
</rdf:Description>
<rdf:Description rdf:about="http://www.geoinfo.uji.es/data/datasetGML.gml">
  <ore:isAggregatedBy rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
    http://www.geoinfo.uji.es/resource/aggregation
  </ore:isAggregatedBy>
</rdf:Description>
<rdf:Description rdf:about="http://www.geoinfo.uji.es/process/Kml2Gml">
  <ore:isAggregatedBy rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
    http://www.geoinfo.uji.es/resource/aggregation
  </ore:isAggregatedBy>
</rdf:Description>
<rdf:Description rdf:about="http://www.geoinfo.uji.es/proxy/r?
  what=http://www.geoinfo.uji.es/data/datasetKML.kml&
  where=http://www.geoinfo.uji.es/resource/aggregation">
  <ore:proxyIn rdf:resource="http://www.geoinfo.uji.es/resource/aggregation"/>
  <ore:proxyFor rdf:resource="http://www.geoinfo.uji.es/data/datasetKML.kml"/>
  <ores:InputFor rdf:resource="http://www.geoinfo.uji.es/proxy/r?
    what=http://www.geoinfo.uji.es/process/Kml2Gml&
    where=http://www.geoinfo.uji.es/resource/aggregation"/>
  <ores:InputAggregatedBy rdf:resource="http://www.geoinfo.uji.es/resource/
    aggregation"/>
</rdf:Description>
<rdf:Description rdf:about="http://www.geoinfo.uji.es/proxy/r?
  what=http://www.geoinfo.uji.es/process/Kml2Gml&
  where=http://www.geoinfo.uji.es/resource/aggregation">
  <ore:proxyIn rdf:resource="http://www.geoinfo.uji.es/resource/aggregation"/>
  <ore:proxyFor rdf:resource="http://www.geoinfo.uji.es/process/Kml2Gml"/>
<ores:hasInput rdf:resource="http://www.geoinfo.uji.es/proxy/r?
  what=http://www.geoinfo.uji.es/data/datasetKML.kml&
  where=http://www.geoinfo.uji.es/resource/aggregation"/>
<ores:hasOutput rdf:resource="http://www.geoinfo.uji.es/proxy/r?
  what=http://www.geoinfo.uji.es/data/datasetGML.gml&
  where=http://www.geoinfo.uji.es/resource/aggregation"/>
</rdf:Description>
<rdf:Description rdf:about="http://www.geoinfo.uji.es/proxy/r?
  what=http://www.geoinfo.uji.es/data/datasetGML.gml&
  where=http://www.geoinfo.uji.es/resource/aggregation">
  <ore:proxyIn rdf:resource="http://www.geoinfo.uji.es/resource/aggregation"/>
  <ore:proxyFor rdf:resource="http://www.geoinfo.uji.es/data/datasetGML.gml"/>
  <ores:OutputFor rdf:datatype="http://www.geoinfo.uji.es/proxy/r?
    what=http://www.geoinfo.uji.es/process/Kml2Gml&
    where=http://www.geoinfo.uji.es/resource/aggregation"/>
  <ores:OutputAggregatedBy rdf:resource="http://www.geoinfo.uji.es/resource/
    aggregation"/>
</rdf:Description>
```

Figure 4. RDF/XML based representation for a geoprocessing service description

over the AR entities, such as the order in which the services should be organized and what resources acts as input and output parameters.

As observed in Figure 5, the ores:next and ores:previous relationships introduce the partial order among the resources AR-2 and AR-4 of the collection. Also, the ores:hasInput,

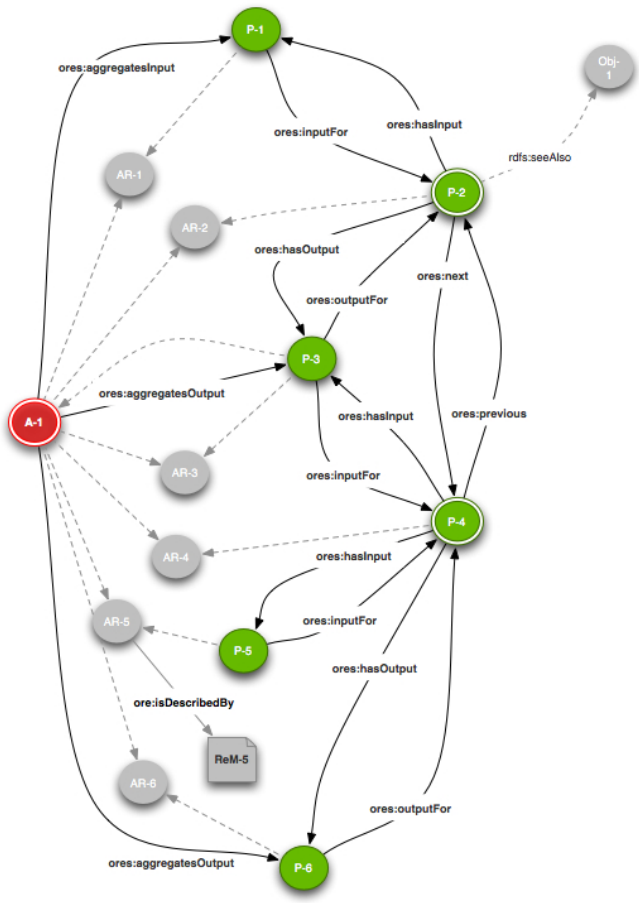



Figure 5. Interlinking geoprocessing services in OAI-ORE

ores:hasOutput and their inverse relationships connect properly all of the aggregated resources. Note that the Proxy entity P-3 has two relations –ores:inputFor and ores:outputFor–, but each one refers to disjoint Proxy entities (P-2 and P-4 respectively). This means that the entity P-3 plays a different role depending on which target entity is connected to.

In this scenario the geoprocessing service referenced by the resource AR-4 and its Proxy entity P-4 represents an intersection function that accepts two inputs: a GML geometry returned by the Kml2Gml service and a second geometry defined by the resource AR-5. The latter represents in fact another collection as indicated by the relationship ore:isDescribedBy that links the Aggregated Resource entity with the Resource Map entity for serialization. As indicated in previous sections, the taks of nesting and reusing aggregations involves the use of OAI-ORE built-in relationships along with the proposed ores:aggregatesOutput (and/or its inverse relationship ores:outputAggregatedBy when required) for indicating those resources that can be reused in other collections. By using these relationships all the geometries required to perform the intersection function through

the capacity resource AR-4 are indicated conveniently in the collection AR-5. Similarly the different outputs generated by the execution of the interlinked geoprocessing services in our example are anotated by using these relationships. It is important to note that not only the final result is indicated (intersection function result) but also intermediate results (Kml conversion to Gml).

Connecting to alternative resources is also possible and even desirable. The P-2's outcomming relationship rdfs:seeAlso lets us connect to other functional-like services for data transformation. In practice, therefore, composing and reusing entire aggregations and aggregated resources by "aggregation by linking" mechanism is possible and even encouraged in the specification of the OAI-ORE protocol itself [22].

## IV. Related work

Geospatial data have been traditionally disposed in collections. Raster data files covering the same area normally come as sets of files that form a collection, and in consequence metadata descriptions are also organized in nested collections [33] [34]. This imposes some degree of linkage since data are grouped according to a certain geographic criterion (proximity, overlay, etc.). The term collection here embraces resources and metadata together, no longer stored separately. Besides, as highlighted throughout this paper, we emphasize tremendous heterogeneity with regard to resources [35], say, geoprocessing services, multimedia resources, and raster data may be part of the same collection.

Regarding geoprocessing services, several research works deal directly with the OGC WPS specification [5] [7] [36]. In most cases, composing properly geospatial services remains still unsatisfactory due to the complexity inherent in some service specification (eg. WFS interface). Other alternatives to distributed geoprocessing computing are semantic-based [37], grid-enabled [38], and REST extensions for BPEL [39].

In the digital libraries domain, recent works have explored the connections between OAI-PMH and LOD communities [40]. OAI-ORE related development has been constantly increasing since its birth, not only at server but also at desktop level appearing different tools that allow users to create their own collections or compound objets. This is the case of LORE [41] a tool created for authoring and publishing compound objects or collections based on the OAI-ORE model for representing bibliographic relationships.

Recent works reveal an increasing interest in connecting geospatial resources of any type [42]. In the SDI community, Florczyk et al. (2010) are exploiting semantic linkages to services in SDIs [43]. The authors propose a linked ontology of administrative units for referencing the same geographic concept to the corresponding instances form multiple WFS services. Similar examples come from the Ordnance Survey, with the publication of the Administrative Geography of Great Britain, an initiative to publish administrative units as linked data sources [44]. Shade and Cox (2010) have recently pointed the similarities between LOD and SDI since geospatial data encoded in GML permit simple mappings to RDF [45]. Our approach built on the OAI-ORE data model goes in this line to provide better support for metadata of individual resources and entire collections. Each resource and its metadata form a logical unit no longer separated, which enables greatly the discovery, access, and linkage to resources and collections.

## V. Conclusion and future work

This paper has presented an ongoing approach to conciliate geospatial services and data with external LOD datasets. The use and extension of OAI-ORE protocol to model collections of interlinked geoprocessing services allows users to regroup and restructure the spectrum of data and services over the Web, in order to build functional, cross-domain Web applications. What is novel here is the straightforward, direct method for describing and packaging resources and collections, compared with the family of business process languages (BPEL, etc.), which makes it easy to browse, compose, access and visualize collections of interlinked geospatial resources.

Applications consuming properly structured data are still missing [26]. In this sense, our future research efforts are centred on building suitable tools to support the creation of OAI-ORE collections and their visualization over virtual globe platforms. Other challenging tasks concern with connecting current SDI catalogue services and applications (geoportals, etc.) to LOD datasets in order to link the relatively small SDI community to others much bigger, so that geospatial researchers may tackle multidisciplinary projects that expand the boundaries of geospatial information.

## Acknowledgment

This work has been partially supported by the CENIT "España Virtual" project funded by the CDTI in the program "Ingenio 2010" through *Centro Nacional de Información Geográfica* (CNIG), and the EuroGEOSS project (ref. 226487) under the EU FP7 programme.

## References

[1] C. Granell, C. Abargues, L. Díaz, and J. Huerta, "Interlinking geoprocessing services," in *Proc. of the IEEE Intl Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2010)*, IEEE Press, 2010, pp. 99-104.

[2] M.F. Goodchild, "Citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0," *Intl. J. of Spatial Data Infrastructures Research*, vol. 2, pp. 24-32, 2007.

[3] M. Craglia, "Next Generation Challenges," in *Proc. of the AGILE Intl. Conference on Geographic Information Science (AGILE 2009)*, Hannover, Germany, Jun. 2009.

[4] M. Craglia *et al.*, "Next-Generation Digital Earth: A position paper from the Vespucci Initiative for the Advancement of Geographic Information Science," *Intl. J. of Spatial Data Infrastructures Research,* vol. 3, pp. 146-167, 2008.

[5] T. Foerster, B. Schaeffer, J. Brauner, and S. Jirka, "Integrating OGC Web Processing Services into Geospatial Mass-Market Applications," in *Proc. IEEE Intl. Conference on Advanced Geographic Information Systems & Web Services (GEOWS 2009)*, IEEE Press, 2009, pp. 98-103.

[6] T. Foerster, B. Schaeffer, S. Jirka, and J. Brauner, "Integrating Web-based Sensor Information into Geospatial Mass-market Applications through OGC Web Processing Services," *Intl. J. on Advances in Intelligent Systems*, vol. 2, no. 2&3, pp. 278-287, 2009.

[7] C. Granell, L. Díaz, and M. Gould, "Service-oriented applications for environmental models: reusable geospatial services," *Environmental Modelling & Software*, vol. 25, no. 2, pp. 182-198, 2010.

[8] J. Brauner, T. Foerster, B. Schaeffer, and B. Baranski, "Towards a Research Agenda for Geoprocessing Services," in *Proc. of the AGILE Intl. Conference on Geographic Information Science (AGILE 2009)*, Hannover, Germany, Jun. 2009.

[9] C. Bizer, R. Cyganiak, and T. Heath, "How to Publish Linked Data on the Web," Tech. Rep., 2007. [Online] Available: http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/

[10] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data - The Story So Far," *Intl. J. on Semantic Web and Information Systems*, vol. 5, no. 3, pp. 1-22, 2009.

[11] M.F. Goodchild, M. Egenhofer, R. Fegeas, and C. Kottman, *Interoperating Geographic Information Systems*. Norwell, MA: Kluwer Academic Publishers, 1999.

[12] P. Zhao, G. Yu, and L. Di, "Geospatial Web Services," in B.N. Hilton (Ed.), *Emerging Spatial Information Systems and Applications*, pp. 1-35. Hershey, PA: IDEA Group, 2007.

[13] Open Geospatial Consortium Inc., "Web Map Service Implementation Specification 1.3.0," 2006. [Online] Available: http://www.opengeospatial.org/standards/wms

[14] Open Geospatial Consortium Inc.,"Web Feature Service Implementation Specification 1.1.0," 2005. [Online] Available: http://www.opengeospatial.org/standards/wfs

[15] Open Geospatial Consortium Inc.,"Web Coverage Service Implementation Specification 1.1.2," 2008. [Online] Available: http://www.opengeospatial.org/standards/wcs

[16] Open Geospatial Consortium Inc.,"OpenGIS Catalogue Service Implementation Specification 2.0.2," 2007. [Online] Available: http://www.opengeospatial.org/standards/specifications/catalog

[17] L. Bernard, I. Kanellopoulos, A. Annoni, and P. Smits,"The European Geoportal – one step towards the establishment of a European spatial data infrastructure," *Computers, Environment and Urban Systems*, vol. 29, no. 1, pp. 15-31, 2005.

[18] C. Granell, L. Díaz, and M. Gould, "Distributed Geospatial Processing Services," in M. Khosrow-Pour (Ed), *Encyclopedia of Information Science and Technology, Second Edition*, pp. 1186-1193. Hershey, PA: Information Science Reference, 2008.

[19] Open Geospatial Consortium Inc.,"OpenGIS Web Processing Service 1.0.0," 2007. [Online] Available: http://www.opengeospatial.org/standards/wps

[20] N. Alameh, "Chaining Geographic Information Web Services," *IEEE Internet Computing*, vol. 7, no. 5, pp. 22-29, 2003.

[21] K. Holtman and A. Mutz, "Transparent Content Negotiation in HTTP," Internet Engineering Task Force (IETF) Memo RFC 2295. [Online] Available: http://www.ietf.org/rfc/rfc2295.txt

[22] Open Archives Initiative, "Open Reuse and Exchange," 2008. [Online] Available: http://www.openarchives.org/ore/

[23] Open Archives Initiative, "ORE User Guide - Abstract Data Model," Oct 2008. [Online] Available: http://www.openarchives.org/ore/1.0/datamodel

[24] Open Archives Initiative, "Protocol for Metadata Harvesting," 2002. [Online] Available: http://www.openarchives.org/pmh/

[25] A. Maslov, A. Mikeal, S. Phillips, J. Leggett, and M. McFarland, "Adding OAI-ORE Support to Repository Platforms," in *Proc. of the 4th Intl. Conference on Open Repositories (OR'09)*, Atlanta, Georgia, 2009.

[26] M. Hausenblas, "Exploiting Linked Data to Build Web Applications," *IEEE Internet Computing*, vol. 13, no.4, pp. 68-73, 2009.

[27] M. Nottingham and R. Sayre, "The Atom Syndication Format," Internet Engineering Task Force (IETF) Memo RFC 4287. [Online] Available: http://www.ietf.org/rfc/rfc4287.txt

[28] C. Granell, M. Gould, R. Gronmo, and D. Skogan, "Improving Reuse of Web Service Compositions," in *Proc. of the 6th Intl. Conference on E-Commerce and Web Technologies (EC-Web 2005)*, LNCS, 2005 pp. 358-367.

[29] Open Geospatial Consortium Inc.,"OGC KML 2.2," 2008. [Online] Available: http://www.opengeospatial.org/standards/kml/

[30] Open Geospatial Consortium Inc.,"OpenGIS Geography Markup Language (GML) Encoding Standard 3.2.1," 2007. [Online] Available: http://www.opengeospatial.org/standards/gml

[31] M. Lutz, "Ontology-based descriptions for semantic discovery and composition of geoprocessing services," *GeoInformatica*, vol. 11, no. 1, pp. 1-36, 2007.

[32] S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D.L. McGuinness, P.F. Patel-Schneider, L.A. Stein, and F.W. Olin, "OWL Web Ontology Language Reference," W3C Recommendation, 2004. [Online] Available: http://www.w3.org/TR/owl-ref/

[33] M.F. Goodchild and J. Zhou, "Finding geographic information: Collection-level metadata," *GeoInformatica*, vol. 7, no. 2, pp. 95-112, 2003.

[34] J. Nogueras-Iso, F.J. Zarazaga-Soria, and P.R. Muro-Medrano, *Geographic information metadata for spatial data infrastructures: resources, interoperability and information retrieval*. Berlin: Springer, 2005.

[35] Y. Raimond, C. Sutton, and M. Sandler, "Interlinking Music-Related Data on the Web," *IEEE MultiMedia*, vol. 16, no. 2, pp. 52-63, 2009.

[36] A. Friis-Christensen, R. Lucchi, M. Lut, and N. Ostlander, "Service chaining architectures for applications implementing distributed geographic information processing," *Intl. J. of Geographical Information Science*, vol. 23, no. 5, pp. 561-580, 2009.

[37] P .Yue, L. Di, W. Yang, G. Yu, and P. Zhao, "Semantics-based automatic composition of geospatial web service chains," *Computers & Geosciences*, vol. 33, no. 5, pp. 649-665, 2007.

[38] A. Chen, L. Di, Y. Wei, Y. Bai, and Y. Liu. "Use of grid computing for modeling virtual geospatial products," *Intl. J. of Geographical Information Science*, vol. 23, no. 5, pp. 581-604, 2009.

[39] C. Pautasso, "RESTful Web service composition with BPEL for REST," *Data & Knowledge Engineering*, vol. 68, no. 9, pp. 851-866, 2009.

[40] B. Haslhofer and B.Schandl, "Interweaving OAI-PMH Data Sources with the Linked Data Cloud," *Intl. J. Metadata, Semantics and Ontologies*, vol. 5, no.1, pp. 17-31, 2010.

[41] A. Gerber and J. Hunter, "A Compound Object Authoring and Publishing Tool for Literary Scholars based on the IFLA-FRBR," *Intl. J. of Digital Curation*, vol. 4, no. 2, pp. 28-42, 2009.

[42] M. Gahegan, J. Luo, S.D. Weaver, W. Pike, and T. Banchuen, "Connecting GEON: Making sense of the myriad resources, researchers and concepts that comprise a geoscience cyberinfrastructure," *Computers & Geosciences*, vol. 35, no. 4, pp. 836-854, 2009.

[43] A. Florczyk, F.J. Lopez-Pellicer, R. Bjar, J. Nogueras-Iso, and F.J. Zarazaga-Soria, "Applying Semantic Linkage in the Geospatial Web," in *Proc. of the AGILE Intl. Conference on Geographic Information Science (AGILE 2010)*, Guimaraes, Portugal, May 2010.

[44] J. Goodwin, C. Dolbear, and G. Hart, "Geographical Linked Data: The Administrative Geography of Great Britain on the Semantic Web," *Transactions in GIS*, vol. 12, no. s1, pp. 19-30, 2008.

[45] S. Schade and S. Cox, "Linked Data in SDI or How GML Is Not about Trees," in *Proc. of the AGILE Intl. Conference on Geographic Information Science (AGILE 2010)*, Guimaraes, Portugal, May 2010.

# An Ontological Framework for Autonomous Systems Modelling

Julita Bermejo-Alonso*, Ricardo Sanz†, Manuel Rodríguez‡ and Carlos Hernández§
*Autonomous System Laboratory (ASLab)*
*Universidad Politécnica de Madrid, Madrid, Spain*
* *Email: jbermejo@etsii.upm.es*
† *Email: ricardo.sanz@upm.es*
‡ *Email: manuel.rodriguezh@upm.es*
§ *Email: carlos.hernandez@upm.es*

***Abstract*—Conceptual modelling aims at identifying, and characterising the entities and the relationships of a selected phenomenon in some domain. The obtained conceptual models express the meaning of the concepts used by domain experts, and the relationships between them. An ontology is a formal specification of a common conceptualisation shared by stakeholders and experts in a domain. Ontologies can serve as the semantic support for conceptual modelling, guiding and constraining the intended meaning of the conceptual models. We have followed this approach in our model-based control systems, by developing a domain ontology and an ontology-based methodology to support the conceptual modelling of autonomous systems. The ontology for autonomous systems captures the ontological elements to describe an autonomous system's structure, function, and behaviour. The methodology exploits the ontology to generate the conceptual models for a generic engineering process. Both elements have been used in case studies to obtain conclusions on the suitability of the developed ontology and its application in the model-based engineering of autonomous systems.**

***Keywords*-ontology-based engineering; autonomous systems; ontology-driven conceptual modelling.**

## I. INTRODUCTION

Our research addresses the development of a universal technology for autonomous systems, in a model-based control paradigm, where the autonomous system uses models as its main knowledge representation supporting the control system's operation. The first stage is to develop these conceptual models based on the design knowledge used by developers to describe and to engineer them. The next step is for the autonomous system itself to use these models, to decide about the actions to be taken. The aim is to make the autonomous system to operate with the same models the engineers use to build it, to be used by cognitive control loops in the autonomous system to increase their autonomy.

We have followed an ontological approach to support autonomous systems conceptual modelling and engineering [1]. The developed framework consists of two intertwined elements: a domain ontology and an ontology-based methodology. The ontology for autonomous systems (OASys) captures the ontological elements to describe an autonomous system's structure, function, and behaviour. The OASys-based methodology exploits OASys to generate conceptual models for a generic engineering process. Both elements have been used in case studies, as to obtain conclusions on the suitability of the developed ontology and its application in the model-based engineering of autonomous systems.

The paper is structured as follows. Section II describes the Autonomous System (ASys) research programme under development for a technology of autonomous systems. Section III reviews related work on ontologies for autonomous systems. Section IV provides the description of the framework: an ontology for autonomous systems, as well as the methodology which exploits the ontological elements. Section V presents the case studies and their models, explaining how the conceptual models have been developed following the ontology-based methodology. Section VI provides some concluding remarks, as well as suggesting further work to address in next stages of our research.

## II. THE AUTONOMOUS SYSTEM RESEARCH PROGRAMME

The context is the long-term research project ASys, which addresses the development of a technology for the systematic development of autonomous systems. This includes methods for analysing requirements, architectures of autonomy and reusable assets with a cross-domain approach.

During their operation, systems might need to put up with external perturbations, changes from the original specification, or unexpected dynamics not always predicted. Production and automation engineers desire autonomous systems, capable of working on its own. However, this autonomy is considered to be bounded, i.e., the system to be fully autonomous but constrained by the engineers who have developed it. This is, in a sense, a strong point of difference between natural and artificial autonomy. Natural autonomous systems are considered to be truly autonomous systems in the etymological sense of the word (i.e., following their own behavioural laws). On the other side, artificial autonomous systems shall behave autonomously but only to a certain extent, being bounded by externally imposed restrictions (e.g., concerning safety, economics or environmental impact).

The strategical decision in the ASys project is the consideration of all the domain of autonomy, i.e., to cater for any

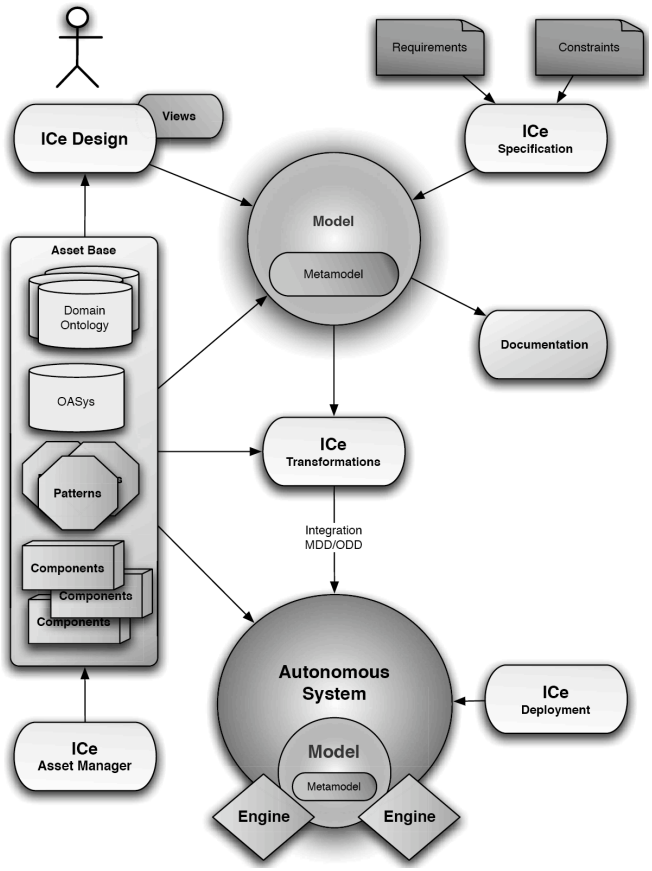



Figure 1. Elements in the ASys Research Programme

autonomous system regardless of its particular application. This implies a wide range of (autonomous) systems to be considered in the research, from robot-based applications to continuous processes or pure software systems. While this movement towards generality may imply a consubstantial reduction of powerfulness of the developments (abstraction may convey vacuity) the progressive domain focalisation [2] method will enable the derivation of progressively domain-focused assets that are, at the same time, compliant with the abstractions and capable of providing functional value. The strategy followed to increase a system's autonomy will be by exploiting a particular class of patterns: *cognitive control loops*, which are control loops based on knowledge [3].

The research programme considers different elements to materialise the former ideas (Figure 1): an architecture-centric design approach, a methodology to engineer autonomous systems based on models, and an asset base of modular elements to fill in the roles specified in the architectural patterns.

Our engineering process covers from the initial specifications and knowledge, to the final product, i.e., the autonomous system. The first stage of the research programme focuses on ontologies, as a common conceptualisation to describe domain knowledge. Both a survey of existing domain ontologies and the development of an ontology for the domain of autonomous systems (OASys) [4] are addressed. One of the central goals is to produce a methodology to exploit these ontologies to generate models based on the knowledge they contain.

A cornerstone of the ASys programme is the use of design patterns as the core vehicle for reusable architecture exploitation [5]. Design patterns present solutions to recurring design problems in a certain context. ASys patterns could be classified in two categories: *architectural patterns*, that express the structural organization of an autonomous system, i.e., they realise the architecture, and *domain patterns*, that describe a mechanism to solve a concrete but recurring problem in a particular context, in a similar way to Buschmann's categorisation into architectural patterns and design patterns [6]. Patterns will not be used in ASys independently from the OASys ontology. Domain patterns will describe interactions of the system's components and with the environment, by using the conceptualisation of the ontology, that represent design solutions so that the behaviour of the system fulfills the engineering requirements. They will model the intended - designed - system's dynamics with its environment. On the other hand, architectural patterns will do the same for the internal system's organization and dynamics, describing them in terms of interactions in between the ontological elements that conceptualize the system itself. Thus all system patterns will not only be specified departing from the OASys concepts, but eventually will become part of the ontology itself, modelling the relations and interactions between them as designed by the engineers.

Models constitute the core for our autonomous systems research programme. The type and use of models to be specifically developed for the autonomous system are initially considered. User and designer requirements, and constraints imposed by the system itself will guide the development of the models. The ASys Model Development Methodologies will address this model characterisation and development. The next stage is to extract from the built models a particular view of interest for the autonomous system. Unified functional and structural views are considered critical for our research as they provide knowledge about the intentions and the behaviours of the autonomous system.

The obtained models will be exploited by means of commercial application engines or customised model execution modules. Considering the metacognitive needs of autonomous systems, metamodels are addressed in the research. Progressive domain focalisation will be used to address the different levels of abstraction between metamodels and models. As an additional product documentation about the built models for their update, exchange, and query will be obtained.

The ASys programme approach is conceptual and architecture-centric. The suitability of extant control and

cognitive control architectures shall be determined, in terms of how they match the ASys research ideas and developed products (ontologies, models, views, engines). Possible adaptations and extensions of the analysed architectures shall also be considered.

The definition and the consolidation of architectural patterns that capture ways of organising components in functional subsystems will be critical. As a generalisation strategy, the different elements considered in the ASys research programme will be assessed in different testbeds.

## III. RELATED WORK

An ontology is a formal specification of a common conceptualisation shared by stakeholders and experts in a domain [7] [8]. Two dimensions can be considered in any ontology: the *pragmatic dimension* to express its application to a certain domain with an intended use, developed following a specific methodology or design method; and the *semantic dimension* as a representation mechanism that structures, organises and formalises a particular content, according to a level of granularity.

An ontology contains classes, relationships, instances and axioms. Classes correspond to entities in the domain under analysis. Relationships relate such entities. Instances are the actual objects that are in the domain. Axioms constrain the use of all former elements.

Focusing on autonomous systems, ontologies are used as a knowledge representation mechanism, in terms of a specification of a conceptualisation. Hence, the ontology contains the concepts to be manipulated by the actors of the autonomous system. The ontological elements are defined based on a computational language, such as OWL or UML.

A first example of ontologies developed for autonomous systems relates to their use for mobile robots. The underlying idea is to conceptualise the different entities taking part in the operation of a mobile robot. Ontologies describe the environment, as a repository of the objects in it [9] or the location the robots are moving in [10] [11]. Mobile robots also need to construct and manipulate representations of the surrounding environment built up by means of sensors, where ontologies represent spatial knowledge [12] [13] [14].

Mobile robots are usually faced with real-time and complex tasks which might require extremely large knowledge to be stored and accessed. Ontologies help to structure this knowledge and its different levels of abstraction [15], to describe task-oriented concepts [16], to act as metaknowledge for learning methods and heuristics [17] or to define concepts related to actions, actors and policies to constraint behaviour [18].

Additionally, ontologies have been used for agent-based systems, where the stress has been on using ontologies for knowledge sharing and exchange among the different agents in the system [19]. Common, global or shared ontologies are used to overcome the semantic heterogeneity among agents. Commitments to the shared ontology permit the agents to interoperate and cooperate while maintaining their autonomy [20] [21] [22] [23] [24].

Recently, ontologies have been used within autonomic computing developments, as a knowledge representation mechanism to provide support for information exchange and integration. Autonomic systems require knowledge from different sources to be represented in a common way. The shared conceptualisation allows to reduce structural and semantic heterogeneity, which appears when the autonomic system handles data whitin different schemes, contents or intended meanings. Moreover, having ontology axioms and rules to verify the model play a major role, when autonomic systems are faced to a high heterogeneity of data and semantics [25] [26] [27]. As for other autonomous systems, ontologies have been developed to the describe the autonomic system [28] or its environment [29].

Autonomous systems have thus benefited of using ontologies for their design and operation [30] [31] :

- Ontologies clarify the structure of knowledge: performing an ontological analysis of a domain allows to define an effective vocabulary, assumptions and the underlying conceptualization. The analysis also allows to separate domain knowledge from operational or problem-solving one.
- Ontologies help in knowledge scalability: knowledge analysis can result in large knowledge bases. Ontologies help to encode and manage them in a scalable way.
- Ontologies allow knowledge sharing and reuse: by associating terms with concepts and relationships in the ontology as well as a syntax for encoding knowledge in them, ontologies allow further users to share and reuse such knowledge.
- Ontologies increase the robustness: ontological relationships and commitments can be used to reason about novel or unforeseen events in the domain.
- Ontologies provide a foundation for interoperability among heterogeneous agents and system's elements.

Hence, ontologies provide a common conceptualisation that can be shared by all those involved in an engineering development process. They also procure a good mean to analyse the knowledge domain, allowing the separation of descriptive and problem-solving knowledge. They can be as generic as needed allowing its reuse and easy extension. Ontologies can serve as the semantic support for conceptual modelling, guiding and constraining the intended meaning of the conceptual models of the autonomous system.

## IV. THE OASYS FRAMEWORK

The OASys Framework captures and exploits the concepts to support the description and the engineering process of any autonomous system. This has been done by developing two different elements:

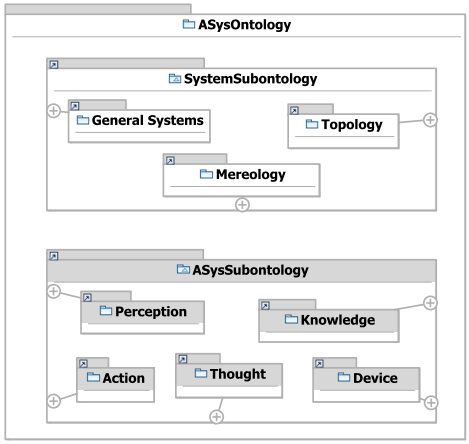



Figure 2. ASys Ontology structure

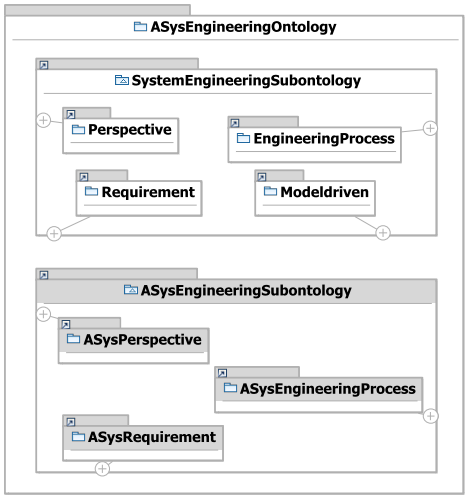



Figure 3. ASys Engineering Ontology structure

- An autonomous systems domain ontology (OASys): The term autonomous systems covers too broad a domain, being applied to a wide range of systems. Finding common features and elements shared by different autonomous systems will aid for their description and engineering. Hence, a domain ontology describing concepts, relationships and axioms related to the autonomous systems domain, as a representation-based mechanism built on UML (a tool suitable to represent ontologies as well as to be used for model-driven engineering to develop the final software assets). The aim is to provide generic building elements to cater from generic autonomous systems to specific applications.
- An OASy-based Engineering Methodology: Once the domain ontology is described, it serves as the basis for an ontology-based engineering process for autonomous systems, whose structure aims at being re-usable and scalable.

Software-intensive systems are usually developed in an ad-hoc engineering process, i.e., specifically built to fulfill specific requirements of a project or application, without general guidelines which will allow a further partial or full re-use. Furthermore, a myriad of different users (either human or artificial) interact in the engineering process, generally without sharing a common conceptualization of the problem. The development process is usually prone to misunderstandings, as well as time- and effort-consuming. A common conceptualisation of relevant concepts and development techniques will be useful to aid in the requirements specification and development phases of an autonomous system engineering process.

### A. The Ontology for Autonomous Systems: OASys

OASys has been developed to describe the domain of autonomous systems, as software and semantic support for the conceptual modelling of autonomous system's description and engineering. The pragmatic dimension of the ontology, i.e., its application to the domain of autonomous systems, has been addressed considering two different ontologies as part of it: the ASys Ontology to conceptualise the autonomous system's description, and the ASys Engineering Ontology to do likewise with the autonomous system's engineering process.

For its development, we have followed the METHONTOLOGY methodology [32] that provides the guidelines and steps for ontological engineering, from knowledge acquisition to ontology evaluation.

As starting point, the *knowledge acquisition phase* where documents and ontologies have been reviewed. Documents have been analysed to come up with existing terminology and definitions for the different domains, subdomains, applications and aspects considered in the ontology's structure. The sources included articles, technical reports describing body of knowledge, and books. As underlying focus, the research and ideas developed in the ASys research programme. Existing well-established glossaries and ontologies have also been assessed to be integrated in the ontology.

Next, the *development activities* to obtain a prototype of the ontology which evolves as new versions are considered. These activities encompass, among others, the specification, the conceptualisation, and the formalisation of the ontology:

- The specification activity aims at answering questions such as which domain is considered, which use is given to the ontology, and who will use it.
- The conceptualisation activity organises and structures the knowledge acquired using external representations that are independent of the knowledge representation and implementation paradigms in which the ontology will be formalised and implemented afterwards.
- The formalisation activity has as goal to formalise the

Table I
SYSTEM SUBONTOLOGY PACKAGES

| Package | Purpose |
|---|---|
| General Systems | To provide concepts to characterise any kind of system's structure, function and behaviour |
| Mereology | To gather general concepts for whole-part relationships |
| Topology | To collect general concepts for topological connections |

Table II
ASYS SUBONTOLOGY PACKAGES

| Package | Purpose |
|---|---|
| Perception | To define concepts to describe the perceptive and sensing process in an autonomous system |
| Knowledge | To specify concepts to describe the different kinds of knowledge (models, ontologies, goals) used by an autonomous system |
| Thought | To characterise concepts to describe the goal-oriented processes in an autonomous system |
| Action | To collect concepts about the operations carried out and performing actors in an an autonomous system |
| Device | To gather concepts to describe the autonomous system's devices |

Table III
SYSTEM ENGINEERING SUBONTOLOGY PACKAGES

| Package | Purpose |
|---|---|
| Requirement | To gather concepts to describe systems requirements |
| Perspective | To provide concepts to describe different perspectives in a system |
| Engineering Process | To define concepts to describe the engineering process |
| Model-driven | To collect concepts to describe model-driven engineering |

Table IV
ASYS ENGINEERING SUBONTOLOGY PACKAGES

| Package | Purpose |
|---|---|
| ASys Requirement | To specify concepts to describe the requirements in an autonomous system |
| ASys Perspective | To gather concepts to describe an autonomous system from different aspects |
| ASys Engineering Process | To provide concepts to describe the costruction process of an autonomous system |

conceptual model. This has been made using UML [33] to implement the ontological elements in the ontology.

To consider the semantic dimension of the ontology, the contents of the ASys and ASys Engineering Ontologies have been organised and structured into subontologies and packages. Subontologies are used to organise the ontological elements into generic knowledge to domain-specific ones. Each subontology contains different packages that gather aspect-related concepts and relationships. Both subontologies and packages have been designed to allow future extensions and updates to OASys.

The ***ASys Ontology*** contains the concepts, relations, attributes and axioms to characterise an autonomous system, organised in two subontologies (Figure 2):

- The System Subontology contains the elements necessary to define any systems structure, behaviour and function, based on general and well-established theories, which consists of different packages (Table I).
- The ASys Subontology elements specialise the System Subontology concepts and relationships for an autonomous system. The subontology is organised in different packages which address the different aspects and functionalities in an autonomous system: the perception, thought, and action activities based on knowledge that an autonomous system performs (Table II).

The ***ASys Engineering Ontology*** collects the ontological elements to describe and support the construction process of an autonomous system. It comprises two subontologies to address at a different level of abstraction the ontological elements to describe the autonomous system's engineering development process (Figure 3).

- The System Engineering Subontology gathers concepts and relationships related to process and software system's engineering. It is based on different metamodels, specifications and glossaries which have been long used for software-based developments. The subontology contains different packages to address different aspects to consider, as needed, within a system's development: requirements, engineering process, perspective, and model driven (Table III).
- The ASys System Engineering Subontology contains the specialisation of the System Engineering Ontology contents, as well as additional ontological elements to describe a concrete autonomous system's engineering process, organised as different packages (Table IV). To mention that this subontology does not yet include an ASys Model-driven package, since the specific model-driven development for autonomous systems in the ASys research project will be later specified.

### *B. The OASys-based Engineering Methodology*

The OASys-based Engineering Methodology is a generic ontology-based, autonomous systems engineering method based on OASys ontological elements. The methodology focuses on the description on how to carry out the autonomous system generic engineering process, having as guideline the ontological elements in the System Engineering and ASys System Engineering Ontologies. Being OASys-based, the methodology considers the ontological elements required in

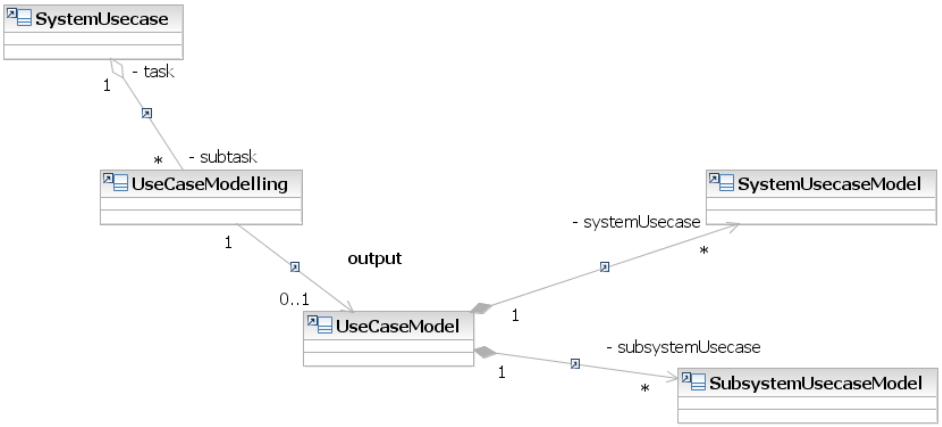



Figure 4. The Use Case Modelling subtask


Figure 5. The Use Case Detailing subtask

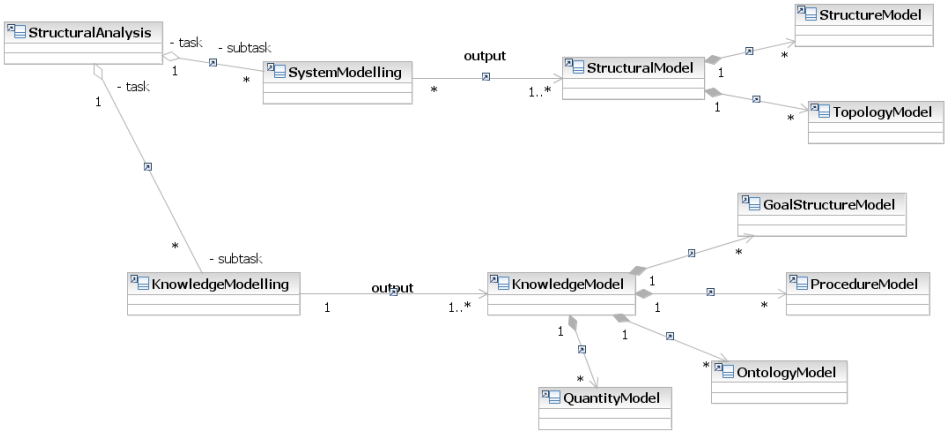



Figure 6. The Structural Analysis task

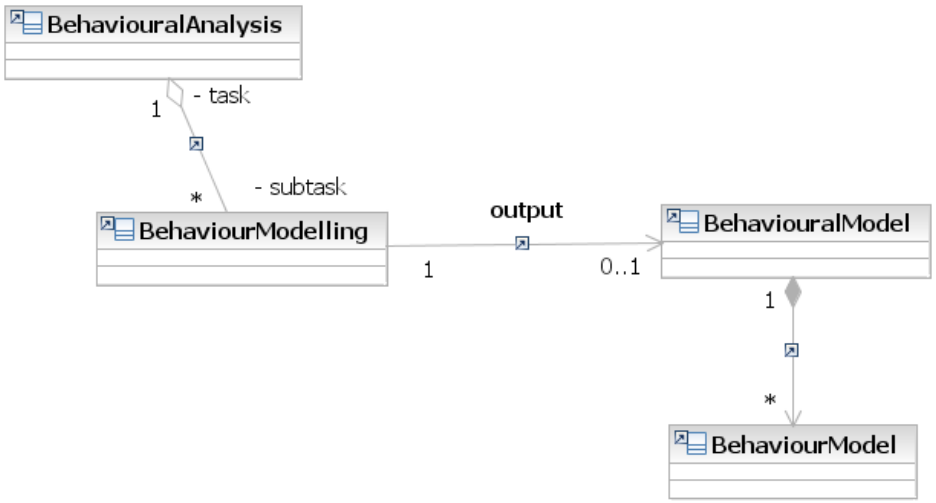



Figure 7. The Behavioural Analysis task

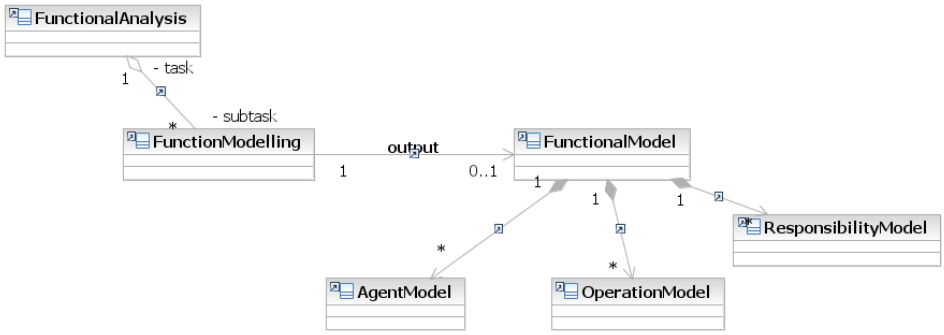



Figure 8. The Functional Analysis task

the different tasks, by specifying the OASys packages to be used.

The engineering methodology tries to transfer state-of-the-art systems engineering methods to the lame engineering practices in cognitive autonomous systems construction. The methodology consists of two main phases: ASys Requirement to identify the autonomous system requirements, and ASys Analysis to consider the autonomous system analysis of its structure, behaviour and function.

The *ASys Requirement* phase identifies and elicits stakeholders' requirements for the system, to characterise the processes and the system. Traditional requirements engineering techniques are used to specify the requirements. The System Use Case task obtains the system and subsystem's use cases models through the Use Case Modelling subtask (Figure 4) and the Use Case Detailing subtask (Figure 5). All the elements shown in these UML class diagrams are defined as ontological elements in the ASys Engineering Ontology.

The different models obtained as workproducts of this phase are constructed using the ontological elements defined in the Requirement and ASys Requirement packages, as it is exemplified in Section V.

The *ASys Analysis* phase describes the autonomous system from different viewpoints, paying attention to the structural, behavioural and functional features in the autonomous system. The concepts and relationships considered to describe the tasks, subtasks and workproducts are defined as ontological elements in the ASys Engineering Ontology.

The Structural Analysis task (Figure 6) considers the system from a structural viewpoint, consisting of different modelling subtasks to analyse the system's subsystems and elements, variables, algorithms, and ontologies. The main work products obtained are the Structural and the Knowledge Models. The Structural Model focuses on modelling the mereotopological relations among the subsystems and elements in the autonomous system, by refining the General Systems, Mereology, and Topology packages. The Knowledge Model considers the different kinds of knowledge for an autonomous system, specialising the Knowledge Package concepts into different models.

The Behavioural Analysis task (Figure 7) targets the system from a behavioural viewpoint, obtaining a Behavioural Model through a Behaviour Modelling subtask, which captures autonomous system behaviour. The Behavioural Model is obtained by specialising the ontological elements in the General Systems Package.

The Functional Analysis task (Figure 8) analyses the system from a functional viewpoint, obtaining a Functional Model that specialises the concepts from the Perception, Thought and Action packages. The Functional Model consists of different models: the Agent Model identifies the different actors in the autonomous system, as defined in the Action Package; the Operation Model captures the operations performed, as specialisation of the Operation taxonomy in the Action Package and the Perception ones defined in the Perception Package; and the Responsibility Model specifies the distribution of responsibilities among the different actors, identifying their responsibilities, possible inputs, outputs and resources to perform an operation.

## V. Testbed Applications

OASys and its related methodology have been applied in the description and engineering of two testbeds considered within the ASys project: the Robot Control Testbed (RCT), and the Process Control Testbed (PCT). The OASys-based Engineering Methodology has been applied to the testbeds during their analysis. Different tasks have been carried out and several work products have been obtained. The next sections exemplify some of the conceptual models obtained for the testbeds. The terms in italics refer to those ontological elements defined in the different Subontologies and Packages of OASys. Relationships between concepts do not follow this format, for the sake of simplicity.

### A. The Robot Control Testbed

The Robot Control Testbed (RCT) is a collection of mobile robot systems, with a wide range of implementations and capabilities (from conventional Simultaneous Localisation And Mapping (SLAM) based mobile robots to virtual robots inspired in rat brain neuroscience). The research aim is to develop a general, customisable autonomous robot architecture providing the system with the necessary cognitive capabilities and robustness to perform complex tasks in uncertain environments.

The current robotic application, Higgs, consists of different interrelated systems to provide the desired capabilities for autonomous navigation (Figure 9). Three different subsystems, composed of several elements, can be considered: the base platform, the onboard systems, and the supporting systems.

The base platform is a mobile robot ActivMedia Pioneer 2-AT8. It is a robust platform, specially designed for outdoor applications. The platform includes all the necessary elements to implement a control and navigating system.

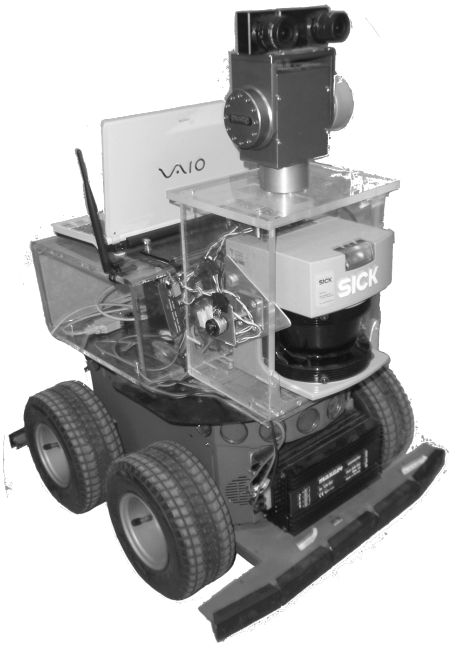



Figure 9. Base platform with onboard systems (Higgs)

Several element have been added to the base platform as onboard systems. Hardware systems have been attached to the platform to expand the range of functionalities, such as an onboard computer, lasers, cameras and a GPS. Moreover, software modules have been developed to provide the mobile robot with further capabilities such as synthetic emotions, communication, surface recognition and real-time processes.

Additionally, other supporting systems have been included to complete the testbed, such as a wireless network, servers and remote controllers to aid in the control operations of the mobile robot.

### B. RCT Conceptual Modelling

This section describes the conceptual models obtained for the current robotic-based system, as a result of applying the OASys-based Engineering Methodology for the RCT *Requirement Phase*. In this *Phase*, the *Requirement*s are identified, considering the *RequirementViewpoint* concept defined in the Requirement Package. *Requirement*s are *organised* by means of *UseCase*s as defined in the Requirement Package. *UseCase*s are modelled in the *Use Case Modelling Subtask* as part of the *System UseCase Task*, as defined in the ASys Engineering Process Package.

To analyse the *UseCase*s, the Requirement Package' ontological elements in the System Engineering Subontology are used as support. A *UseCase* in OASys has been defined as a mean to capture a requirement of a system, as defined in the Requirement Package. To define the UseCase, the *Subject* as system under consideration, and the different *UseCaseActor*s

as objects that interact with the system are also identified, among other aspects. As a result, a *System UseCase Model* is obtained, detailing the previous identified elements. The UML classes in the model are instantiations of the original OASys concepts, this fact being specified by the UML roles names in the shown associations.

When the former ontological concepts are particularised for the Robot Control Testbed, an RCT *UseCase Model* is obtained, which shows the RCT's requirements by means of use cases. A system's requirements can be of different types (physical, functional, performance, interface, design) as defined in the Requirement Package. An initial requirement analysis made by the RCT developers, identified the *FunctionalRequirement*s for the RCT. A FunctionalRequirement is defined in OASys as a requirement that specifies an operation or behaviour that a system must perform. Primary *FunctionalRequirement*s for the RCT are to navigate, as well as to survive.

The navigation *Requirement* is captured by means of the *UseCase* Navigation, which *include*s the secondary *FunctionalRequirement*s of being able to explore the environment, to identify elements in the environment, and to avoid obstacles. These requirements are captured in the *UseCase*s of EnvironmentExploration, Identification and ObstacleAvoidance respectively (Figure 10). In turn, the *FunctionalRequirement* of surviving is captured in the Survival *UseCase*, which *include*s the SubsystemFailure and Recharge *UseCase*s (Figure 11).

It was found interesting to detail a particular *UseCase* by paying attention to the *Subsystem*s in the *System*, to detail further *Requirement*s. *Subsystem*s identified in the RCT are the BasePlatform, the OnboardSystem, and the SupportingSystem. Focusing, for example, on the Navigation *UseCase* previously defined, it is possible to identify additional *Requirement*s for the *Subsystem*s, in the form of a RCT Subsystem *UseCaseModel* (Figure 12).

### C. The Process Control Testbed

The second case study to which the ontological framework is being applied is the Process Control Testbed (PCT), which is the chemical pilot plant designed to test the application of autonomous system ideas to continuous processes. Its main component is a Continuous Stirred Tank Reactor (CSTR), as well as the related instrumentation and control systems. The aim is to provide the plant with cognitive capabilities to carry out complex tasks such as fault diagnosis, alarm management, and control system reconfiguration.

The current system is shown in Figure 13. It is composed of a CSTR type reactor (R01) where two input streams are fed by their corresponding pumps (P01 and P02), and an outlet stream. The chemical reactions taking place in the reactor should be cooled with water or heated with steam, depending on the reaction. The stirring unit assures the homogeneous composition in the reactor. There is also a

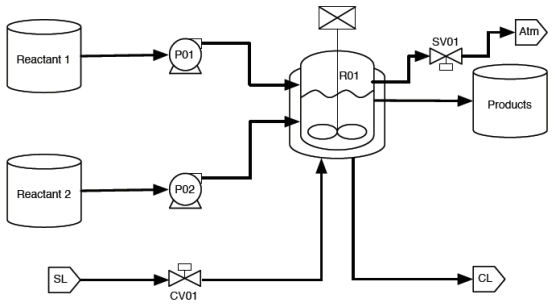



Figure 13. PCT Process Diagram

relief valve (SV01). The hydraulic subsystem is composed of the different tubing (for the reactants, the products, and the cooling water and steam), the pumps used to feed the reactants to the reactor, and the control (CV1) and relief (SV01) valves. Two small tanks for reactants storage and one for the product are also part of the system.

The instrumentation subsystem consists of several sensors (one for pH, one for temperature, and one for pressure). Moreover, there is an oxidation-reduction potential (ORP) sensor, an electrochemical analyzer, and two electromagnetic flowmeters to test and to control the different elements (reactants, products and steam) in the system. The control system is composed of a conventional computer, together with National Instrument's data acquisition boards for input/output analog and digital signals (current and voltage).

### D. PCT Conceptual Modelling

The examples of applying the OASys-based Engineering Methodology here focus on the *Structural Analysis* for the PCT, considering the *StructureViewpoint*. This *Viewpoint* pays attention to the PCT structure, as considering the *ASysStructure Concern*. A PCTStructuralViewpointModel can be obtained (Figure 14) by instantiating the different concepts, relationships and attributes in the Perspective and the ASysPerspective Packages. The original concepts have been ontologically instantiated into PCT related ones.

The *Structural Analysis* has as objective the analysis of a system considering its structural aspects, under a *StructuralViewpoint*. Different *Engineering Model*s in the form of *Structural Model*s can be obtained as result of performing the *System Modelling Subtask* defined in the ASys Engineering Package. The *Structural Model* is a model kind to describe an autonomous system from a structural viewpoint that conforms, as a matter of fact, to a specific level of detail that could be further refined.

For the PCT, the *Structural Model* consists of different *Structure Model*s to describe the elements, as well as the *Topology Model*s to describe topological connections among the components in the system. The Structure Model specialises the ontological elements in the General Systems and Mereology Packages, to describe the structural features of

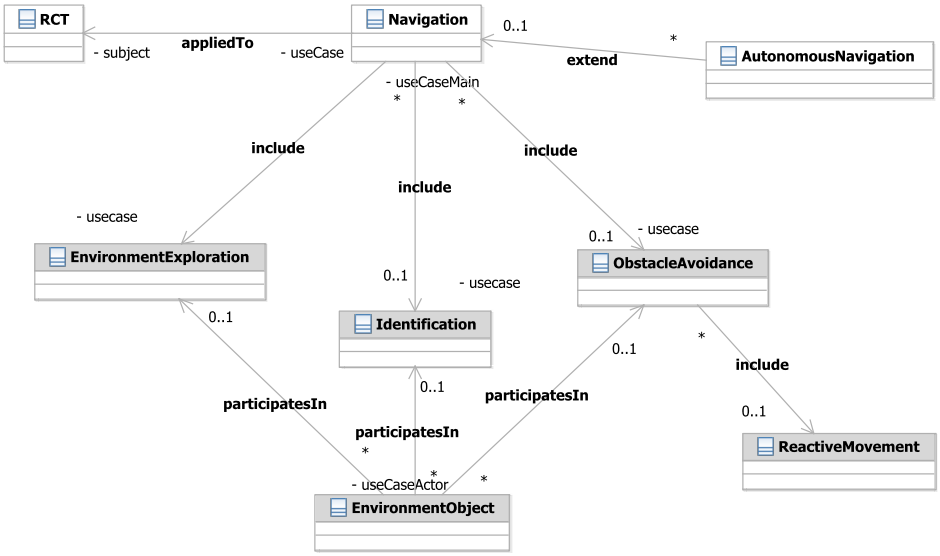



Figure 10. RCT Navigation UseCase Model

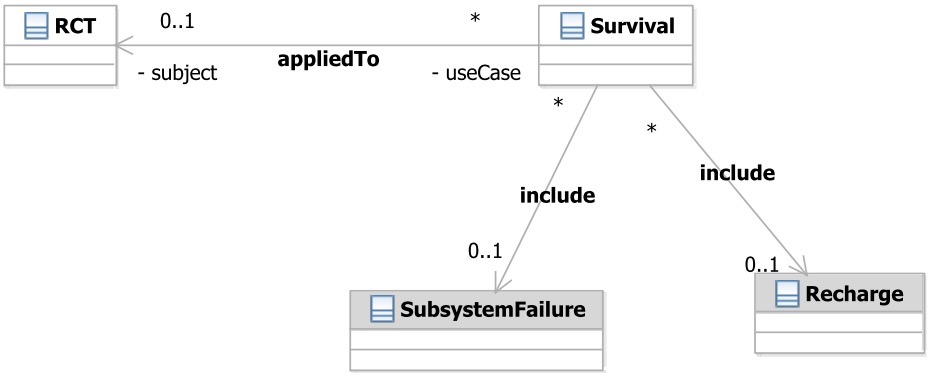



Figure 11. RCT Survival UseCase Model

the testbed. The Topology Model, not shown here, instantiates the concepts in the Topology Package.

The PCT Structure Model shows that in general, a *System* consists of *Subsystems*, i.e., a system that is constituent of another one, which in turn can be decomposed until reaching to its *Elements*, which cannot be further decomposed, as defined in the GST Package. The Process Control Testbed consists of the CSTR Reactor, the Hydraulics subsystem, the Control System, the Instrumentation, and the Power (Figure 15).

In the model, the relationships between the *System* (the PCT), and the *Subsystem*s (CSTRreactor, Hydraulics, ControlSystem and Power) are expressed by means of the UML composition association. To point out that the relationships between the different *Subsystems* and *Elements* are modelled directly using UML composition, aggregation and generalization associations, detailing the classes roles in them. If required, the associations could be changed into directed associations detailed by using the terms in the Mereology Package. However, it has not been done here to avoid cluttering the model.

Additional *Structure Model*s were obtained when a particular *Subsystem* such as the CSTR Reactor (Figure 16), the Hydraulics (Figure 17), or the Instrumentation (Figure 18) were analysed.

The combined use of these ontological concepts provides for the description of the structural relations among the components in the Process Control Testbed.

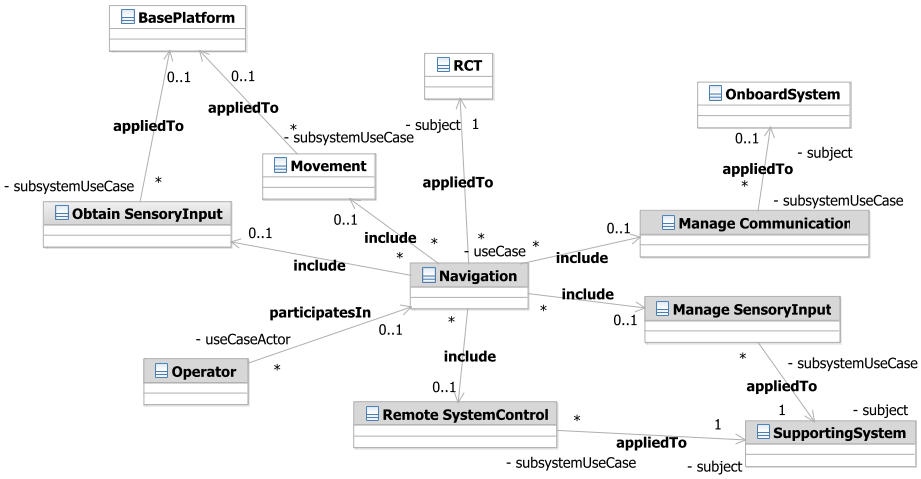



Figure 12. RCT Navigation UseCase Model detailed for Subsystems

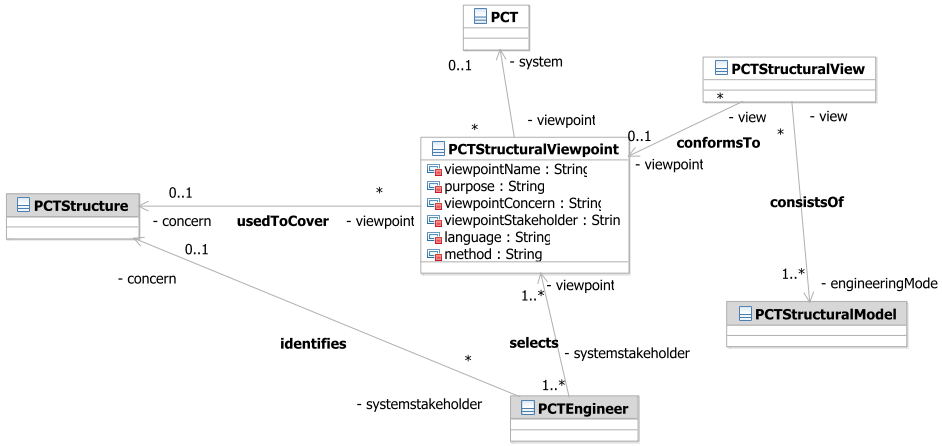



Figure 14. PCT StructuralViewpoint Model

## VI. Conclusion and Further Work

Our research has considered the autonomous systems domain with a global view. Previous ontologies developed for this kind of systems focused on a constrained or limited approach, either being mobile robots or agent based systems. The approach here has been to define the different elements to describe the system structure and function in a way general enough to be reused among different applications. The resulting ontology is intended to be generically applicable to the engineering of varied systems, being this is reflected in its structure in terms of subontologies and packages.

Engineers developing autonomous systems describe and characterise them considering different elements that take part in their operation: the perception process, the knowledge to be used, the system's goals, the functional decomposition, as well as the actors to carry out the system's actions. OASys has catered for all these aspects, by means of the subontologies and packages that have gathered, conceptualised and formalised the ontological elements related to each one of them. OASys has come to cover the problem of modelling in a modular and unified view the complex systemic structures that many autonomous systems exhibit.

Its structure and content have allowed to scale and manage the broad range of concepts, and prevent imprecise definitions and mismatches when referring to autonomous systems. The separation between the description and the engineering aspects in a system, by means of the ASys Ontology and the ASys Engineering Ontology, has allowed to address independently the characterisation of the testbeds.

Moreover, the existence of different levels of abstraction within the ontology has shown its suitability to describe an autonomous system using a domain focalisation technique. OASys can be further complemented with additional subdomain, task or application ontologies, without losing its reusability and generality features.

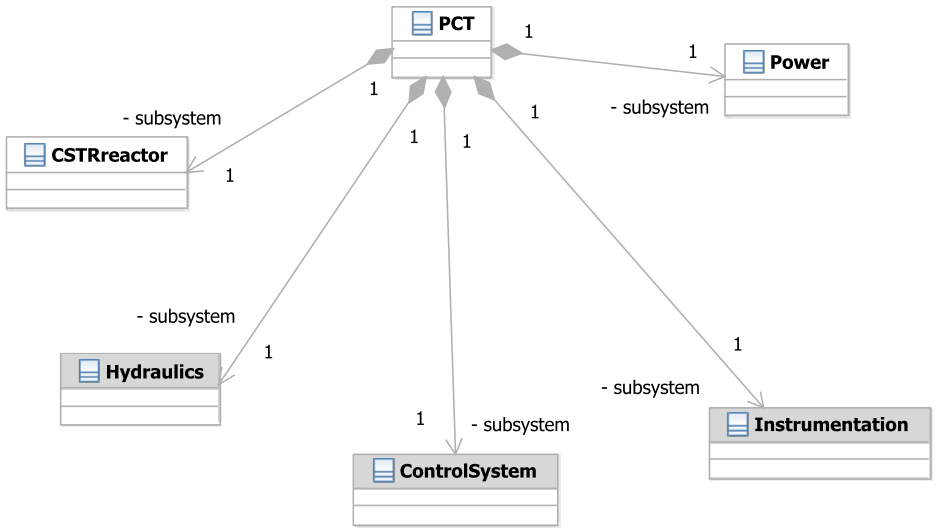



Figure 15. PCT Structure Model

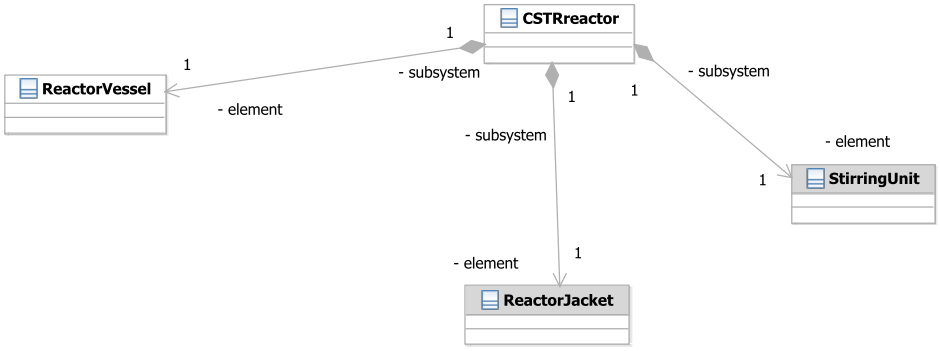



Figure 16. CSTR Reactor Structure Model

The ontology for autonomous systems has also served as an ontological metamodel to support the OASys-based Engineering Methodology. This methodology has suggested how autonomous systems can be characterised and should be engineered using conceptual models. The methodology has provided a practical approach to reuse and to extend the ontology for concrete applications. Viewpoints provide an adequate concept to tackle the complexity that is typical to this kind of systems.

The engineering methodology has provided a way of systematically devising an ontology-based account of systems going down to functional aspects that end up with the algorithmic design.

The ontology-based conceptual modelling and engineering process offers improvements on the knowledge sharing and reuse between the applications developers. OASys and its related methodology have helped to increase the realiability of the software models to be obtained. There are no previous attempts to engineer autonomous systems ontologies and an associated methodology for its application.

The OASys-based Engineering Methodology benefits from the underlying ontological commitments and relationships to build up the autonomous system's different conceptual and engineering models, ensuring against traditional meaning and conceptual mismatches during its development.

Our aim is for any autonomous system to use the conceptual models based upon the ontological framework described in this paper, as part of a model-based systems engineering strategy [34].

An autonomous system is to perform using models of its environment, of its actions and of itself to operate, as in the model-based control paradigm, where those models will be the same ones used by the engineers to build the system. This will ultimately provide the system with self-engineering

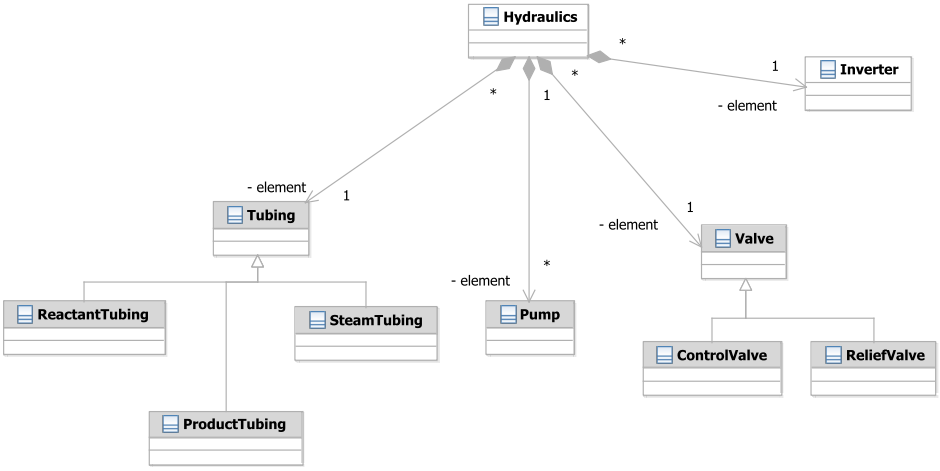



Figure 17. Hydraulics Structure Model

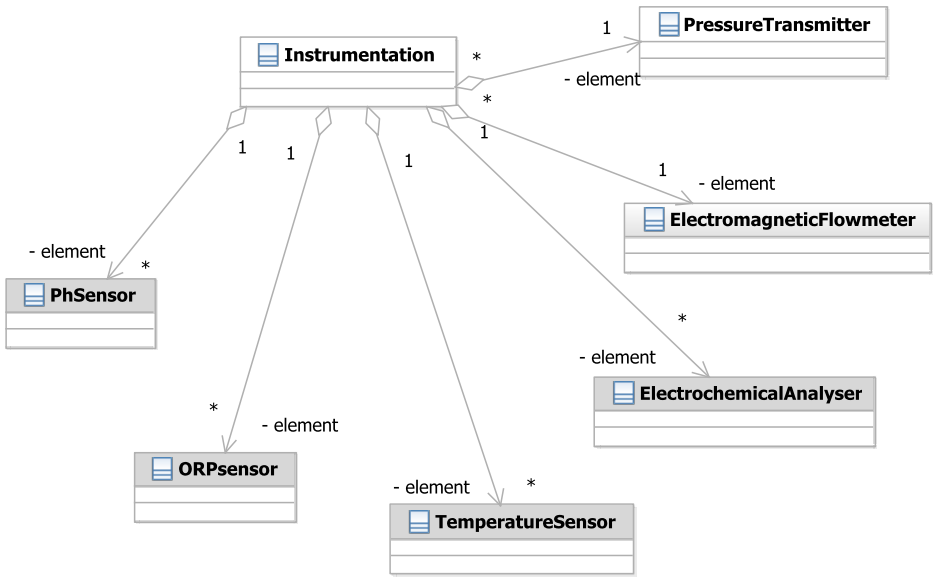



Figure 18. Instrumentation Structure Model

capabilities required for robust autonomy [35]. A further step will be for the agents to interpret those models based on OASys. It is yet to investigate how OASys and these models obtained by instantiating its ontological elements, can be used to generate meaning for the decision-making process of the different actors of an autonomous system [36].

Next stages envision the development of a methodology based on ontological and software patterns with the aid of conceptual modelling tools. This refinement process of concepts to address higher level of detail, as well as the ontological elements usage is to be defined in the ASys Modelling Methodology, as a next step in the overall ASys research programme. This methodology will define, among other aspects, how a concept is selected, how to integrate a concept into a pattern, how to establish and to import its relationships with other concepts, and how to detail or to add its attributes in the development of a concrete model.

## ACKNOWLEDGMENT

The authors would like to acknowledge the support of the Spanish Ministry of Education and Science (grant C3: Control Consciente Cognitivo) and the European Commission (Grant ICEA: Integrating Cognition, Emotion and Autonomy).

## REFERENCES

[1] J. Bermejo-Alonso, R. Sanz, M. Rodríguez, and C. Hernández, "Ontology–based engineering of autonomous systems," in *Proceedings of the The Sixth International Conference on Autonomic and Autonomous Systems (ICAS 2010)*, M. Bauer, J. L. Mauri, and O. Dini, Eds. Cancun, Mexico: IEEE Computer Society, 7–13 March 2010, pp. 47–51.

[2] R. Sanz, I. Alarcón, I. Segarra, M. de Antonio, and J. Clavijo, "Progressive domain focalization in intelligent control systems," *Control Engineering Practice*, vol. 7, no. 5, pp. 665–671, 1999.

[3] R. Sanz, C. Hernández, and M. Rodriguez, "The epistemic control loop," in *Proceedings of CogSys 2010 - 4th International Conference on Cognitive Systems*, Zurich, Switzerland, January 2010.

[4] J. Bermejo-Alonso, "OASys: ontology for autonomous systems," Ph.D. dissertation, E.T.S.I.I.M., Universidad Politécnica de Madrid, 2010.

[5] R. Sanz and J. Zalewski, "Pattern-based control systems engineering," *IEEE Control Systems Magazine*, vol. 23, no. 3, pp. 43–60, June 2003.

[6] F. Buschmann, R. Meunier, H. Rohnert, P. Sommerland, and M. Stal, *Pattern-oriented Software Architecture*. Wiley, 1996, vol. 1: a system of patterns.

[7] T. Gruber, "Toward principles for the design of ontologies used for knowledge sharing," in *International Workshop on Formal Ontology in Conceptual Analysis and Knowledge Representation*, N. Guarino and R. Poli, Eds. Padova, Italy: Kluwer Academic Publishers, 1993.

[8] R. Studer, V. Benjamins, and D. Fensel, "Knowledge engineering: Principles and methods," *IEEE Transactions on Data and Knowledge Engineering*, vol. 25, no. 1-2, pp. 161–197, 1998.

[9] C. Schlenoff, S. Balakirsky, M. Uschold, R. Provine, and S. Smith, "Using ontologies to aid navigation planning in autonomous vehicles," *The Knowledge Engineering Review*, vol. 18, no. 3, pp. 243–255, 2003.

[10] C. Scrapper and S. Balakirsky, "Knowledge representation for on–road driving," in *Proceedings of the 2004 AAAI Spring Symposium on Knowledge Representation and Ontologies for Autonomous Systems*, Stanford, California, March 2004.

[11] C. Scrapper, S. Balakirsky, and E. Messina, "Self awareness in the mobility open architecture simulation and tools framework," in *Proceedings of the 2005 ACM workshop on Research in knowledge representation for autonomous systems*. ACM Press, 2005, pp. 35–41.

[12] M. Uschold, R. Provine, S. Smith, C. Schlenoff, and S. Balikirsky, "Ontologies for world modeling in autonomous vehicles," in *18Th International Joint Conference on Artificial Intelligence, IJCAI'03*, 2003.

[13] T. Wagner, U. Visser, and O. Herzog, "Egocentric qualitative spatial knowledge representation for physical robots," *Robotics and Autonomous Systems*, vol. 49, pp. 25–42, 2004.

[14] R. Provine, M. Uschold, and S. Smith, "Observations on the use of ontologies for autonomous vehicle navigation planning," in *Proceedings of the 2004 AAAI Spring Symposium on Knowledge Representation and Ontologies for Autonomous Systems*, Stanford, California, March 2004.

[15] T. Barbera, J. Albus, E. Messina, C. Schlenoff, and J. Horst, "How task analysis can be used to derive and organize the knowledge for the control of autonomous vehicles," *Robotics and Autonomous Systems*, vol. 49, pp. 67–78, 2004.

[16] S. Wood, "Representation and purposeful autonomous agents," *Robotics and Autonomous Systems*, vol. 49, pp. 79–90, 2004.

[17] S. L. Epstein, "Metaknowledge for autonomous systems," in *Proceedings of the 2004 AAAI Spring Symposium on Knowledge Representation and Ontologies for Autonomous Systems*, Stanford, Palo Alto, Ca., March 2004.

[18] H. Jung, J. Bradshaw, S. Kulkarni, M. Breedy, L. Bunch, P. Feltovich, R. Jeffers, M. Johnson, J. Lott, N. Suri, W. Taysom, G. Tonti, and A. Uszok, "An ontology–based representation for policy–governed adjustable autonomy," in *Proceedings of the 2004 AAAI Spring Symposium on Knowledge Representation and Ontologies for Autonomous Systems*, Stanford, California, March 2004.

[19] V. Tamma, S. Cranefield, T. Finin, and S. Willmott, Eds., *Ontologies for Agents: Theory and Experiences*, ser. Whitestein Series in Software Agent Technologies and Autonomic Computing. Birkhäuser, 2005.

[20] Q. Chen and U. Dayal, "Multi–agent cooperative transactions for e–commerce," in *Conference on Cooperative Information Systems*, 2000, pp. 311–322.

[21] D. Dou, D. McDermott, and P. Qi, "Ontology translation by ontology merging and automated reasoning," in *Ontologies for Agents: Theory and Experiences*, ser. Whitestein Series in Software Agent Technologies, V. Tamma, S. Cranefield, T. Finnin, and S. Willmott, Eds. Birkhäuser, 2005, pp. 73–94.

[22] A. Malucelli, D. Palzer, and E. Oliveira, "Combining ontologies and agents to help in solving the heterogeneous problem in e-commerce negotiations," in *International Workshop on Data Engineering Issues in E-Commerce (DEEC 2005), IEEE Computer Society*, Tokyo, Japan, April 2005, pp. 26–35.

[23] M. Nodine and J. Fowler, "On the impact of ontological commitments," in *Ontologies for Agents: Theory and Experiences*, ser. Whitestein Series in Software Agent Technologies, V. Tamma, S. Cranefield, T. Finin, and S. Willmott, Eds. Birkhäuser, 2005, pp. 19–42.

[24] V. Tamma, "An ontology model supporting multiple ontologies for knowledge sharing," PhD, University of Liverpool, 2001.

[25] M. Cebulla, "Knowledge-based assessment of behaviour in dynamic environments," in *Proceedings of the 2005 ACM workshop on Research in knowledge representation for autonomous systems*. Bremen, Germany: ACM Press, November 2005, pp. 17–26.

[26] L. Stojanovic, A. Abecker, N. Stojanovic, and R. Studer, "Ontology–based correlation engines," in *Proceedings of the International Conference on Autonomic Computing (ICAC'04)*, I. Computer, Ed., 2004, pp. 304–305.

[27] J. Martin-Serrano, J. Serrat, J. Strassner, G. Cox, R. Carroll, and M. O. Foghlu, "Policy–based context integration and ontologies in autonomic applications to facilitate the information interoperability in NGN," in *Proceedings of the Workshop on Hot Topics in Autonomic Computing*. Jacksonville, FL, U.S.A.: USENIX Association, 2007.

[28] G. Tziallas and B. Theodoulidis, "Building autonomic computing systems based on ontological component models and a controller synthesis algorithm," in *Proceedings of the 14th International Workshop on Database and Expert Systems Applications (DEXA'03)*, Prague, Czech Republic, September 2003, pp. 674–680.

[29] J. C. Strassner, N. Agoulmine, and E. Lehtihet, "FOCALE: a novel autonomic networking architecture," in *Proceedings of the First Latin American Autonomic Computing Symposium (LAACS 2006)*, 2006.

[30] L. Stojanovic, J. Schneider, A. Maedche, S. Libischer, R. Studer, T. Lumpp, A. Abecker, G. Breiter, and J. Dinger, "The role of ontologies in autonomic computing systems," *IBM Systems Journal*, vol. 43, no. 3, pp. 598 – 616, 2004.

[31] B. Chandrasekaran, J. Josephson, and V. Benjamins, "What are ontologies and why do we need them?" *IEEE Intelligent Systems*, vol. 14, no. 1, pp. 20–26, 1999.

[32] M. Fernández-López, A. Gómez-Pérez, and N. Juristo, "METHONTOLOGY: from ontological art towards ontological engineering," in *AAAI'97 Spring Symposium on Ontological Engineering*, A. Farquhar, M. Grüninger, A. Gómez-Pérez, M. Uschold, and P. van der Vet, Eds., Stanford University, CA, U.S.A., 1997, pp. 33–40.

[33] *OMG Unified Modeling Language (OMG UML) Superstructure Version 2.2*, Object Management Group, February 2009.

[34] R. Sanz, C. Hernández, J. Gomez, J. Bermejo-Alonso, M. Rodríguez, A. Hernando, and G. Sanchez, "Systems, models and self–awareness: towards architectural models of consciousness," *International Journal of Machine Consciousness*, vol. 1, no. 2, pp. 255–279, December 2009.

[35] R. Sanz, I. López, and C. Hernández, "Self–awareness in real–time cognitive control architectures," in *Proceedings of the AAAI Fall Symposium on Consciousness and Artificial Intelligence: Theoretical foundations and current approaches*, Washington, D.C., November 2007.

[36] R. Sanz, J. Bermejo, I. López, and J. Gomez, *Toward Artificial Sapience: Principles and Methods for Wise Systems*. Springer London, 2008, ch. A real-time agent system perspective of meaning and sapience, pp. 61–73.

# Motion Planning of Autonomous Agents Situated in Informed Virtual Geographic Environments

Mehdi Mekni
*Department of Computer Science*
*Sherbrooke University*
*Sherbrooke, Canada*
*Email: mmekni@gmail.com*

***Abstract*—Multi-Agent Geo-Simulation (MAGS) aims to simulate phenomena involving a large number of autonomous situated actors (implemented as software agents) evolving and interacting within a Virtual representation of the Geographic Environment (VGE). Motion planning is a critical issue since it corresponds to one of the most important activities of agents moving in a complex and large-scale VGE. There is also a need for an accurate representation of the environment in order to support efficient path planning computation as well as reactive navigation for the detection and avoidance of obstacles and other agents. In this paper, we propose a semantically informed and geometrically precise virtual geographic environment method which allows to use Geographic Information System (GIS) data to automatically build an informed graph structure called Informed Virtual Geographic Environmnt (IVGE). Furthermore, we propose a topologic abstraction algorithm which builds a Hierarchical Topologic Graph (HTG) describing the IVGE and a Hierarchical Path Planning (HPP) algorithm which uses this graph. In addition, we propose a graph-based neighborhood structure in order to support motion planning of autonomous agents taking into account the characteristics of the IVGE.**

***Keywords*-Informed Virtual Geographic Environment (IVGE); Hierarchical Path Planning (HPP); Navigation and Collision Avoidance;**

## I. INTRODUCTION

During the last decade, the Multi-Agent Geo-Simulation (MAGS) approach has attracted a growing interest from researchers and practitioners to simulate phenomena in a variety of domains including traffic simulation, crowd simulation, urban dynamics, and changes of land use and cover, to name a few [3]. Such approaches are used to study phenomena (i.e., car traffic, mobile robots, sensor deployment, crowd behaviours, etc.) involving a large number of simulated actors (implemented as software agents) of various kinds evolving in, and interacting with, an explicit description of the geographic environment called Virtual Geographic Environment (VGE). Nevertheless, simulating such autonomous situated agents remains a particularly difficult issue, since it involves several different research domains: geographic environment modelling, spatial cognition and reasoning, situation-based behaviours, etc. The autonomy of an agent is defined by its capacity to perceive, act and decide about its actions without external governance [23]. One of the most fundamental capacities of a situated autonomous agent is its ability to navigate inside a VGE while taking into account both the agent's and the environment's characteristics. When examining situated agents in a VGE, whether for gaming or simulation purposes, one of the first questions that must be answered is how to represent the world in which agents navigate [25]. Since a geographic environment may be complex and large-scale, the creation of a VGE is difficult and needs large quantities of geometrical data describing the environment characteristics (terrain elevation, location of objects and agents, etc.) [20] as well as semantic information that qualifies space such as buildings, roads, parks, as illustrated in Figure 2. Hence, a situated autonomous agent should consider the semantic information that qualifies the geographic environment in which and with which it interacts. Current approaches usually consider the environment as a monolithic structure, which considerably limits the way that large-scale, real world geographic environments and agent's spatial reasoning capabilities are handled [19].

Path planning is a typical spatial reasoning capability for situated agents in VGE [22]. The problem of path planning in MAGS involving complex and large-scale VGEs has to be solved in real time, often under constraints of limited memory and CPU resources [6]. Classic path planners provide agents with obstacle-free paths between two positions located in the VGE. Such paths do not take into account the environment's characteristics (topologic and semantic) nor different agent categories and capabilities [5]. For example, classic planners assume that all agents are equally capable to reach most areas in a given map, and that any terrain portion which is not traversable by one agent is also not traversable by the other agents. Such assumptions limit the applicability of these planners to solve a very narrow set of problems: path planning of homogeneous agents in a homogeneous environment. A path planning algorithm should take into account the semantic information that qualifies the geographic environment in which agents evolve and with which they interact. Moreover, in navigation applications

(local path planning) which involve several moving agents that do not know their respective mobility plans, a scheme for detection and resolution of collision conflicts between agents becomes mandatory. In this project, our goal is to address the issue of navigation and path planning for agents having different capabilities evolving in complex and large-scale geographic environments.

In order to achieve such a goal, a geographic environment model should precisely represent geographic features. It should also integrate several semantic notions characterising these geographic features. Since we deal with large-scale geographic environments, it would be appreciable to have a VGE organised hierarchically in order to reduce the search space for path planning. Indeed, hierarchical search is recognised as being an effective approach to reduce the complexity of such a problem [10]. There is also a need for autonomous situated agents which are able to plan paths, to detect and avoid both *static* and *dynamic* obstacles located in the VGE. Static obstacles correspond to areas that are not navigable for agents such as walls, fences, trees, rivers, etc. Static obstacles also include obstructions resulting from terrain elevation. Dynamic obstacles correspond to other moving agents which are navigating in the VGE.

In this paper, we propose a novel approach to simulate motion planning of autonomous situated agents in virtual geographic environments. This approach is composed of four parts: 1) a geometrically precise and semantically enhanced virtual geographic environment called *Informed VGE* (IVGE); 2) a topologic abstraction algorithm used to diminish path planning complexity; 3) a *Hierarchical Path Planning* (HPP) algorithm to support motion planning of autonomous agents situated in large-scale geographic environments; and 4) a graph-based structure called *Neighborhood Graph* (NG) to address collisions detection and avoidance between moving agents in 3D virtual environments.

The remainder of this paper starts with a discussion of related works on geographic environment representation using data provided by Geographic Information Systems (GIS) and path planning and navigation in virtual environments. In Section III, we present our approach to automatically create an Informed VGE. Section IV outlines a method to enhance the IVGE description using a topologic abstraction that reduces the size of the topologic graph and enables building a hierarchical topologic graph; Section V presents how we leverage the hierarchical graph structure of the IVGE model in order to support situated reasoning algorithms such as hierarchical path planning. Section VI introduces our model to support navigation in Informed VGE. Finally, we conclude with a discussion and present future works.

## II. Related Works

In this section we provide a brief overview of prior works related to *environment representation*, and *path planning and navigation* in virtual environments.

### A. Environment Representation

Virtual environments and spatial representations have been used in several application domains. For example, Thalmann *et al.* proposed a virtual scene for virtual humans representing a part of a city for graphic animation purposes [8]. Donikian *et al.* proposed a modelling system which is able to produce a multi-level data-base of virtual urban environments devoted to driving simulations [15]. Ali *et al.* used a multi-agent geo-simulation approach to simulate customers' behaviours in shopping malls [1]. More recently, Shao *et al.* proposed a virtual environment representing New York City's Pennsylvania Train Station populated by autonomous virtual pedestrians in order to simulate the movement of people [20]. Paris *et al.* also proposed a virtual environment representing a train station populated by autonomous virtual passengers, in order to characterise the levels of services inside exchange areas [17]. However, since the focus of these approaches is computer animation and virtual reality, the virtual environment usually plays the role of a simple background scene in which agents mainly deal with geometric characteristics. Indeed, the description of the virtual environment is often limited to the geometric level, though it should also contain topological and semantic information for other types of applications. Therefore, most interactions between agents and the environment are usually simple, only permitting to plan a path in a 2D or 3D world with respect to free space and obstacle regions [7].

### B. Path Planning and Navigation

An extensive literature exists on agents' path planning in robot motion planning and virtual environments [13]. Roughly, these methods can be categorised as: *path planning* (global) and *navigation* (local).

*Path Planning:* The path planning issue, which consists of finding an obstacle-free path between two distinct positions located in a VGE, has been extensively studied. The computational effort required to plan a path, using a search algorithm such as A* [16] or Dijkstra [14], increases with the size of the search space [5]. Consequently, path planning on large-scale geographic environments can result in serious performance bottlenecks. However, representing the virtual environment using a hierarchical approach reduces the size of the search space as well as the complexity of path planning algorithms [10]. Two recent hierarchical triangulation-based path planning approaches are described in [6], namely *Triangulation A** and *Triangulation Reduction A**, which

are relevant to our work. *TA** makes use of the *Delaunay Triangulation* (DT) technique to build a polygonal representation of the environment without considering the semantic information. This results in an undirected graph connected by constrained and unconstrained edges, the former being traversable and the latter not. *TRA** is an extension of *TA** and abstracts the triangle mesh into a structure resembling a roadmap. Both TA* and TRA* are able to accurately answer path queries for agents since they make use of the DT technique. However, the abstraction technique used by TA* and TRA* aims at maximising triangle size, which does not reduce the size of the search space. Moreover, both *TA** and *TRA** assume a homogeneous flat environment, which considerably reduces the capacity to handle 3D environments enhanced with semantic information.

*Navigation:* An agent navigation behavior aims at predicting local collisions and avoiding the other navigating agents. Most current models are based on a particle approach proposed by Helbing [11]. However, this model suffers from several shortcomings. First, it cannot predict collisions since it waits for navigating agents to collide before adapting their behavior. In addition, it produces oscillations when adapting directions, which affects the quality of the agents' navigation behavior. Finally, Helbing's model manages very basic agents (particles) and it is difficult to adapt it to more complex simulated actors. Other reactive navigation models exist, including variants of potential fields [9]. They can handle dynamic environments, but suffer from "'local-minima"' problems and may not be able to find a collision-free path, when one exists [13]. Often, these models do not give any kind of guarantee on their behavior. Other navigation algorithms are based on path or roadmap modification, which allows a path to be deformed as a result of obstacle detection. These methods include Elastic Bands [18], Elastic Roadmaps [24], and adaptive roadmaps [21]. Alternatively, Lamarche and Donikian proposed to use a *Neighborhood Graph* (NG) based on a *Delaunay Triangulation* (DT) of the agents' positions filtered by visibility [12]. This structure offers a low computational cost, which enables the simulation of a large number of situated agents [12]. However, this NG does not take into account the terrain's elevation since it is based on a two-dimensional DT. In order to support moving agents in virtual environments, we claim that an NG should take into account terrain elevation. Indeed, a real environment is rarely flat and ignoring this information would distort the neighboring relationship between moving agents located in the virtual environment. In Section VI, we propose an approach which extends Lamarche and Donikian's method and enables us to create a 3D NG to support motion planning of autonomous agents situated in *Informed VGE*.

## III. Computation of IVGE Data

In this section, we present our automated approach to computing the IVGE data directly from vector GIS data. Figure 1) depicts the four stages which compose our approach: *input data selection*, *spatial decomposition*, *maps unification*, and finally the *informed graph generation*.

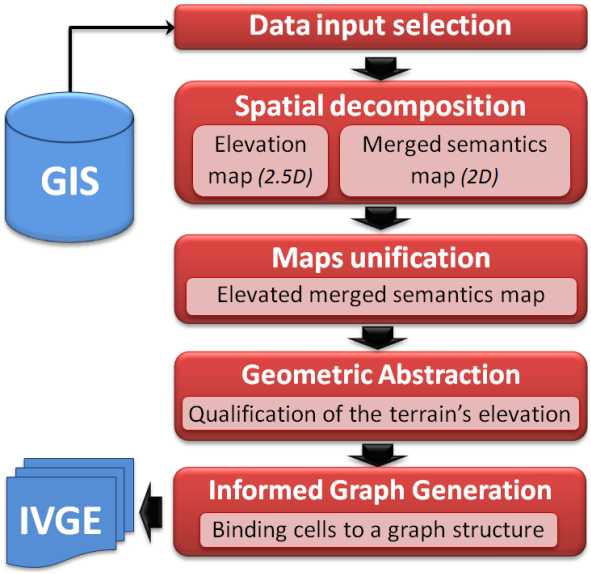
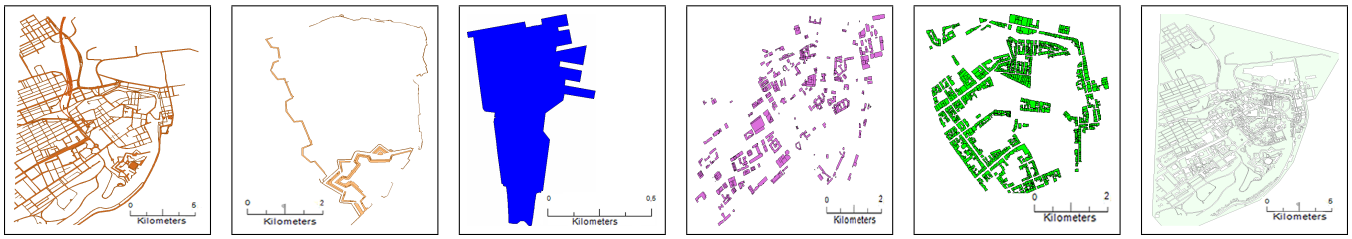



*Figure 1:* The five stages to obtain an IVGE from GIS data.

*Input data selection:* The first step of our approach is the only one requiring human intervention. It consists in selecting the different vector data sets which are used to build the IVGE. The input data can be organised into two categories. First, *elevation layers* containing the geographical marks that indicate absolute terrain elevations. Second, *semantic layers* are used to qualify various types of data in space. Each layer indicates the physical or virtual limits of a given set of features with identical semantics in the geographic environment, such as roads and buildings. The limits can overlap between two layers, and our model can merge the information.

*Spatial decomposition:* The second step consists of obtaining an exact spatial decomposition of the input data into cells. This process is entirely automatic, using a Delaunay Triangulation and can be divided into two parts in relation to the previous phase. First, an elevation map is computed and corresponds to the triangulation of the elevation layer. All the elevation points are injected into a 2D triangulation, the elevation being considered as an additional attribute. This process produces an environment subdivision composed of connected triangles (Figure 3(a)). Such a subdivision provides information about coplanar areas: the elevation of any point inside a triangle can be deduced thanks to the elevation




(a) (b) (c) (d) (e) (f)

*Figure 2:* Various semantic layers related to Quebec city in Canada: (a) road network; (b) old city wall; (c) marina; (d) governmental buildings; (e) houses; (f) sidewalk areas.

of the three original points. Second, a merged semantics map is computed, corresponding to a constrained triangulation of the semantic layers. Indeed, each segment of a semantic layer is injected as a constraint which keeps track of the original semantic data by adding additional attributes. The obtained map is then a constrained triangulation merging all input semantics (Figure 3(b)): each constraint represents as many semantics as the number of input layers containing it.

*Maps unification:* The third step to obtain our IVGE data consists of unifying the two maps previously obtained. This phase can be depicted as the mapping of the $2D$ merged semantic map (Figure 3(b)) onto the $2.5D$ elevation map (Figure 3(a)) in order to obtain the final $2.5D$ elevated merged semantics map (Figure 3(c)). First, preprocessing is carried out on the merged semantics map in order to preserve the elevation precision inside the unified map. Indeed, all the points of the elevation map are injected in the merged semantics triangulation, creating new triangles. Then, a second process elevates the merged semantics map. The elevation of each merged semantics point $P$ is computed by retrieving the corresponding triangle $T$ inside the elevation map, i.e. the triangle whose $2D$ projection contains the coordinates of $P$. Once $T$ is obtained, the elevation is simply computed by projecting $P$ on the plane defined by $T$ using the $Z$ axis.

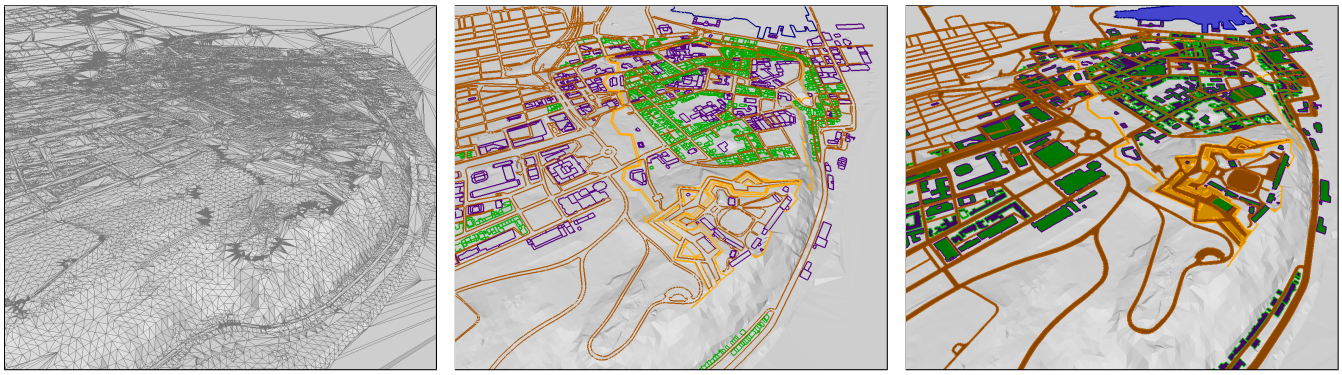
*Informed graph generation:* The resulting unified map now contains all the semantic information of the input layers, along with the elevation information. This map can be used as a topological graph in which each node corresponds to the map's triangles and each arc to the adjacency relations between these triangles. Then, common graph algorithms can be applied to this topological graph, especially graph traversal ones. One of these algorithms retrieves the node, and so the triangle, corresponding to given $2D$ coordinates. Once this node is obtained, it is possible to extract the data corresponding to the position, such as the elevation, and the semantic information. Many other algorithms can be applied, such as path planning and graph abstraction, but they are out of the scope of this paper and will not be detailed here.

## IV. Topologic Abstraction

When dealing with large-scale and complex geographic environments the informed graph becomes very large. The size of a topologic graph has a direct impact on the computation time of the agent's spatial reasoning processes. In order to optimise such a computation time, we need to reduce the size of the informed graph representing the IVGE. The aim of the topologic abstraction is to provide a compact representation of the informed graph suitable for situated reasoning of situated agents. To this end, the topologic abstraction process extends the informed graph with new layers. In each layer (except for the initial layer which is called level 0), a node corresponds to a group of nodes of the immediate lower level. The topologic abstraction simplifies the IVGE description by combining cells (triangles) in order to obtain convex groups of cells. Such a hierarchical structure evolves the concept of *Hierarchical Topologic Graph* (HTG) in which cells are fused in groups and edges are abstracted in boundaries. To do so, convex hulls are computed for every node of the informed graph. Then, the coverage ratio of the convex hull is evaluated as the surface of the hull divided by the actual surface of the node. The topologic abstraction finally performs groupings of a set of connected nodes if and only if the group ratio is close to one. Let $G$ be a group of cells, *Convex* be the convexity rate, and $CH(G)$ be the convex hull of the polygon corresponding to $G$. *Convex* is computed as follows:

$$Convex(G) = \frac{Surface(G)}{Surface(CH(G))} \quad and \quad 0 < Convex(G) \leq 1 \tag{1}$$

Each node $c$ of the informed graph can be topologically qualified according to the number of connected edges given by the $arity(c)$ function: if $arity(c) = 0$ then $c$ is a *closed* cell; if $arity(c) = 1$ then $c$ is a *dead end* cell; if $arity(c) = 2$ then $c$ is a *corridor* cell; and if $arity(c) > 2$ then $c$ is a *crossroads* cell. The topologic abstraction algorithm is based on an in-depth exploration of the informed graph structure. At each step, the algorithm processes cells based on their topology in order to achieve a specific goal: 1) *Virtual Cells*: to characterise the outside of the IVGE; 2)



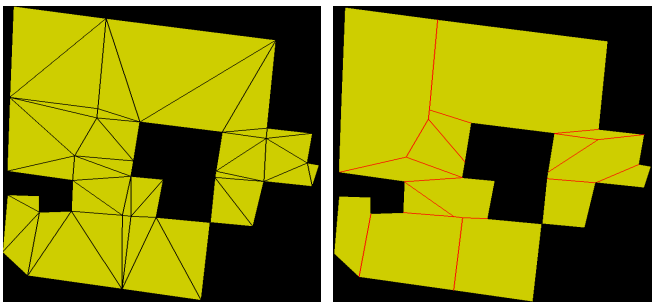
(a) Triangulated elevation map (2.5*D*). (b) Merged semantics map (2*D*). (c) Unified map (2.5*D*).

*Figure 3:* The two processed maps (*a*, *b*) and the unified map (*c*). The semantic colours are the same as in figure 2.

*Access Cells*: to identify access points corresponding to cells connected to at least one virtual cell (Figure ); 3) *Corridor Cells*: to filter excessive discretization of space subdivision in narrow open areas (Figure ); 4) *Crossroads*: to filter excessive discretization of space subdivision due to the misalignment of edges in open areas (Figure ). 5) *Dead End*: Termination of the algorithm.

Let us detail the execution of the topologic abstraction algorithm which starts by processing the *virtual cells* and then their neighboring ones.

- Step 1 (processing of virtual cells): Bring together all the virtual cells in a virtual group, then merge into this group all the adjacent *dead end* cells. Proceed to Step 2.
- Step 2 (processing of access cells): Bring together all the access cells in a single group, then merge into this group all the *access end* cells. Proceed to Step 3, 4, or 5 depending on the type of the neighbooring cell *Cn*.
- Step 3 (processing of corridor cells): If the current cell *Cc* and its neighboor cell *Cn* are of type *corridor*, and if the current group is of type *crossroad*, proceed to Step 3-1, else proceed to Step 3-2.
  - Step 3-1: If $Convex(Gc \cup Cc) > Convex(Cc \cup Cn)$, then merge *Cc* into *Gc* and continue with *Cn*.
  - Step 3-2: Build a new group *Gp* of type *corridor* and assign *Cc* and *Cn* to it.
    - If $Gp = Gn$, where *Gn* corresponds to the group to which belongs *Cn*, then merge *Gc* with *Gp* and *Gn*
    - Else if *Gc* and *Gs* are of type *corridor* and if $Convex(Gc \cup Gp \cup Gn) > \max(Convex(Gc); Convex(Gp); Convex(Gn))$, then merge together *Gc*, *Gp*, and *Gn*.
    - Else if *Gc* and *Gn* are of type *crossroads* and If $Convex(Gp \cup Gn) > \max(Convex(Gp); Convex(Gn))$ then merge together *Gp* and *Gn*.
- Step 4 (processing crossroads cells): Build a group of type *crossroads Gc* and assign the current cell *Cc* to it as well as its neighbooring cells of type *crossroads* or *dead end*. Proceed to Step 3 for cells of type *corridor*.
  - If *Gc* has only one neighboor, turn it into a *dead end* group.
  - Else, if *Gc* has exactly two neighbours, transform it into *corridor* and apply the tests of Step 3;
  - Else, if $Convex(Gc \cup Gn) > \max(Convex(Gc); Convex(Gn))$ for any *Gs* in *dead end* neighboring groups, then merge *Gn* and *Gc*. Repeat the test of Step 4.
- Step 5 (processing dead end cells): Build a new group of type *dead end* and assign the current cell to it. This group will further be merged with its neighbour of type *corridor* or *crossroads* because of the tests in Steps 3 and 4.



(a) (b)

*Figure 4:* Illustration of the topologic abstraction process with a strict convex property (*C(gr) = 1*); (a) the exact space decomposition using CDT techniques (63 triangular cells) ; (b) the topologic abstraction (28 convex polygons)

*Results:* As an illustration, our IVGE generation model has been applied to an urban area representing the center part of Quebec City, with one elevation map and five semantic

layers. The creation of the IVGE takes less than five seconds on a typical computer (Intel Core 2 Duo processor 2.13Ghz, 1Go RAM). The resulting unified map approximately contains 122,000 triangles covering an area of $30km^2$. The necessary time to retrieve the triangle corresponding to a given coordinate is negligible (less than $10^{-4}$ seconds). We applied the topologic abstraction algorithm in order to build a three-level hierarchical and topologic graph. Level 0 corresponds to the informed graph resulting from the exact spatial subdivision. Level 1 of the topologic graph resulting from the topologic abstraction (with soft convex constraint, i.e. $Convex(c) = 1$) reduces the total number of cells (122,000) by merging them into 73,000 convex polygons (called groups) in 2.8 seconds. Level 2 of the topologic graph resulting from the topologic abstraction (with relaxed convex constraint, i.e. $Convex(c) = 0.9$) reduces the total number of groups by merging them into 12,000 convex polygons (called zones) in 1.9 seconds.

## V. Hierarchical Path Planning

In this section, we present our hierarchical path planning algorithm (HPP for short). We then provide a computation analysis of the algorithm complexity which aims to point out the contribution of our algorithm. Finally, we propose a path enhancement method in order to optimise the computed paths for more realistic moving agents.

### A. Algorithm

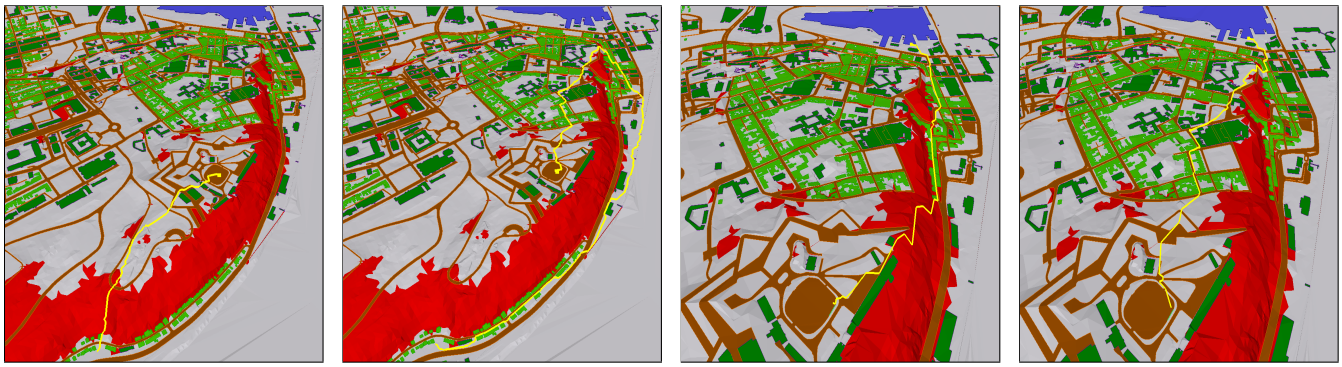
Let us consider the topologic graph extracted from the exact spatial decomposition before highlighting the usefulness of the topologic and semantic abstractions. Since cells are convex, it is possible to build an obstacle-free path by linearly connecting positions located at two different borders belonging to a given cell. Thus, it is also possible to use borders, represented by edges in the graph, to compute obstacle-free paths between different locations in the environment. Since the topologic graph structure is hierarchical, each node at a given level $i$ (except at level 0) represents a group of convex cells or abstract cells of a lower level $i-1$. Hence, our approach can be used to compute a path linking two abstract nodes at any level.

Let us consider a hierarchical topologic graph $G$ composed of $i$ levels. Nodes belonging to level 0 are called *leaves* and represent convex cells produced by the exact spatial decomposition. Nodes belonging to higher levels ($i > 0$) are called *abstract nodes* and are composed of groups. Given a starting position, a final destination, and a hierarchical topologic graph $G$ composed of $i$ levels, the objective of our algorithm is to plan a path from the current position to the destination using $G$. The algorithm starts from the highest level of the hierarchy and proceeds as follows:

- **Step 1:** Identify the abstract nodes to which the starting position and the final destination belong.
  Two cases need to be considered:
  - Case 1: Both are in the same abstract node $k$ at level $i$.
    Proceed to *step 1* with the groups (at level $i-1$) belonging to node $k$.
  - Case 2: They are in different abstract nodes $k$ and $j$ at level $i$. Proceed to *step 2*.
- **Step 2:** Compute the path from the abstract node $k$ to the abstract node $j$.
  For each pair of consecutive nodes ($s$, $t$) belonging to this path, two cases are possible :
  - Case 1: Both are leaves. Proceed to *step 4*.
  - Case 2: Both are abstract nodes. Proceed to *step 3*.
- **Step 3:**
  - If the starting position belongs to $s$ then identify to which group $gs$ of $s$ it belongs and proceed to *step 2*, in order to compute the path from the abstract node $gs$ to the closest common boundary with the abstract node $t$. Else proceed to *step 2* in order to compute the path from the center of the abstract node $s$ to the closest common boundary with the abstract node $t$.
  - If the final destination position belongs to $t$ then identify to which group $gd$ of $t$ it belongs and proceed to *step 2*, in order to compute the path from the closest common boundary with the abstract node $s$ to $gd$. Else proceed to *step 2* in order to compute the path from the closest common boundary with the abstract node $s$ to the centre of the abstract node $t$.
- **Step 4:** Once in a leaf, apply a path planner algorithm (we used the Djkstra and A* algorithms) from the starting position to the final goal using the convex cells which belong to the informed graph.

The strategy adopted in this algorithm is to refine the path planning when getting closer to the destination. The algorithm starts by planning a global path between the start and the destination abstract nodes (step 1). Then, for each pair of successive abstract nodes, it recursively plans paths between groups (of lower levels) until reaching leaves (steps 2 and 3). Once at leaves (convex cells at level 0), the algorithm proceeds by applying a path planning algorithm such as Dijkstra and A* (step 4). Hence, at level $i$, the path planner exploration is constrained by the nodes belonging to the path computed at level $i+1$.

Moving agents can use this algorithm in order to plan paths within the IVGE. The path computed in step 2 is actually a coarse-grained path whose direction is only indicative. Since the path is refined in a *depth-first* way, agents



(a) (b) (c) (d)

*Figure 5:* HPP in the IVGE (the computed path is coloured in yellow). (a) path computed with no regard for the terrain shape; (b) path computed with regard for the terrain shape; (c) and (d) Search paths to get to a place (marina) in the IVGE (place described by semantics) without and with regard to terrain respectively.

can perform local and accurate navigation inside an abstract node without requiring a complete and fine-grained path computation towards the final destination. The lower levels' sub-paths (related to other abstract nodes) are computed only when needed, as the agent moves. Such a *just in time* path planning approach is particularly relevant when dealing with dynamic environments. Classic path planning approaches use the entire set of cells representing the environment and compute the complete path between a start and a final positions. These classical approaches suffer from two major drawbacks : 1) the computation time of a path is considerable since it involves all the cells composing the environment; 2) the planned path may become invalid as a consequence of changes in the environment. An interesting property of our hierarchical path planning approach is the optimization of calculation costs over time. Indeed, the entire path is only computed for the most abstracted graph, which contains a small number of abstract nodes compared to the informed graph (convex cells at level 0). In addition, our approach provides a *just in time* path planning which can accommodate a dynamic environment. Furthermore, this hierarchical path planning is adapted to any type of agents, whenever we are able to generate the abstracted graphs taking into account both the geographic environment and the agents' characteristics.

### B. Complexity Analysis

In order to highlight the outcomes of our approach, let us compare the computation cost of our hierarchical path planning with the standard path planning. Let $G_0(V_0, E_0)$ be the graph representing the virtual environment at level 0, which corresponds to cells produced by the spatial decomposition process. Let $V_0$ correspond to the set of vertices and $E_0$ correspond to the set of edges at level 0. Let $|V_0| = N$ be the number of nodes of the graph $G_0$. Let us consider a starting position $s$ and a destination position $d$ located in the virtual environment. The computation cost of the shortest path between $s$ and $d$ at level 0 (represented by the graph $G_0$) is denoted by $C_0(N)$ and is given by the following equation:

$$C_0(N) = O(N * ln(N)) \tag{2}$$

Let us now compare $C_0(N)$ with the computation cost of our hierarchical path planning algorithm which relies on the hierarchical topologic graph with $k$ levels. To this end, we need to raise some assumptions for the sake of simplification. First, let us assume that the topologic abstraction process may be thought of as a function $h$ which abstracts a topologic graph $G_{i-1}$ and builds a new topologic graph $G_i$. The function $h$ can be written as follows:

$$h(G_{i-1}(V_{i-1}, E_{i-1})) = G_i(V_i, E_i) \quad with \quad 0 \leq i \leq k-1 \tag{3}$$

Let $l_i$ be the *abstraction rate* between two successive levels $i-1$ and $i$ (with $0 \leq i \leq k-1$). Since the abstraction process aims at reducing the number of nodes at each new level, we have $l_i > 1+\varepsilon$ (with $0 \leq i \leq k-1$) as illustrated in equation 4.

$$l_i = \frac{|V_{i-1}|}{|V_i|} \quad with \quad l_i > 1+\varepsilon \; and \; \varepsilon > 0 \tag{4}$$

Second, let us suppose that the $k^{th}$ level of our hierarchical topologic graph is composed of $m$ nodes. $N$ which corresponds to the number of nodes of the graph $G_0$ can be expressed using equations 3 and 4 as follows:

$$N = m * l_{k-1} * ... * l_0 \tag{5}$$

$$N \geq m * (1+\varepsilon)^k \quad with \quad k > 0 \; and \; \varepsilon > 0 \tag{6}$$

$$N = m * \prod_{i=0}^{k-1} (l_i) \quad with \quad k > 0 \; and \; m > 0 \tag{7}$$

Let $l_{Avg}$ be the average value of $l_i$ (with $0 \leq i \leq k-1$). Using $l_{Avg}$, equation 7 becomes:

$$N = m * l_{avg}^k \quad with \quad k > 0 \; and \quad m > 0 \tag{8}$$

Let us replace the term $N$ in equation 2 by its value in equation 8:

$$C_0(m) = O(m * l_{avg}^k * ln(m * l_{avg}^k)) \tag{9}$$

Equation 9 can be developed as follows:

$$C_0(m) = O(m * ln(m) * l_{avg}^k + m * l_{avg}^k * ln(l_{avg}^k)) \tag{10}$$

Let $Nb_k$ be the number of nodes composing the computed path at level $k$. The computation cost of $Nb_k$ is given by the following equation:

$$Nb_k = O(m * ln(m)) \quad with \quad k > 0 \; and \quad m > 0 \tag{11}$$

The hierarchical path planning algorithm involves the computation of the shortest path at level $k$ and the refinement of the path linking each pair of successive nodes at lower levels. Therefore, the shortest path from $s$ to $d$ corresponds to the computation of $Nb_k$ at level $k$ and its refinement through the lowest levels. Such a shortest path is denoted $C_k$ and has a computation cost which can be computed by the following equations:

$$C_k(m) = Nb_k * \sum_{j=0}^{k-1} l_{avg}^j \tag{12}$$

$$C_k(m) = Nb_k * \frac{l_{avg}^k - 1}{l_{avg} - 1} \tag{13}$$

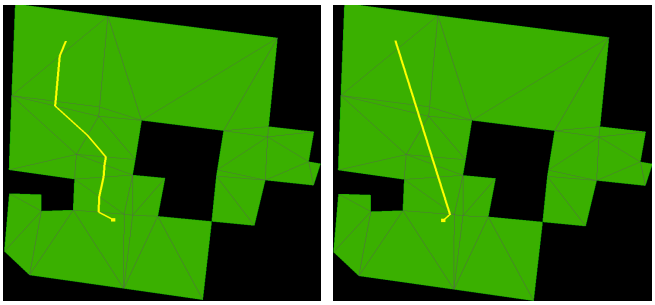
The term $Nb_k$ in equation 13 is replaced by its value expressed in the equation 11 as follows:

$$C_k(m) = O(m * ln(m) + m * \frac{l_{avg}^k - 1}{l_{avg} - 1}) \tag{14}$$

Let us compare the computation costs of standard path planning approaches (equation 10) and our hierarchical path planning approach (equation 14). First, it is obvious that the first term $m * ln(m)$ in equation 10 is inferior to the first term $m * ln(m) * l_{avg}^k$ in equation 14 since the abstraction rate $l_{avg}^k > 1$. Second, in a similar way, the second term $m * (l_{avg}^k - 1/l_{avg} - 1)$ in equation 10 is inferior to the second term $m * l_{avg}^k * ln(l_{avg}^k)$ in equation 14. In conclusion, the hierarchical path planning algorithm along with the hierarchical topologic graph that we propose is at least $ln(l_{avg}^k)$ orders of magnitude faster than standard path planning approaches.

### C. Path Optimisation

The topological abstraction only groups together adjacent cells or groups of cells with respect to the convexity criterion. While this approach is efficient to reduce the size of the topologic graph, it gives up the optimality of the computed path. Indeed, paths are optimal in the abstract graph but not necessarily in the initial problem graph (informed graph at level 0). In order to improve the quality of the computed path (i.e., length and visual optimisation), we perform a post-processing phase called *path optinmisation* (*Figure 6*). Our strategy for path optimisation is simple, but produces good results. The main idea is to replace local sub-optimal parts of the computed paths by straight lines. We start from one end of the path (*Figure 6(a)*) and for each node part of the computed path, we check whether we can reach a subsequent node in the path in a straight line. If this is possible, then the linear path between the two nodes replaces the initial sub-optimal sequence between these nodes (*Figure 6(b)*).



(a) (b)

*Figure 6:* (a) The original computed path ; (b) The computed path after optimisation.

*Results:* The HTG resulting from the topologic abstraction process is particularly suitable to support HPP in IVGE. Two types of HPP have been implemented: 1) a path linking two positions located in the IVGE using the A* algorithm (Figures 5(a) and 5(b)); and 2) a search path linking a position to a qualified area within the IVGE using the *Dijkstra* algorithm (Figures 5(c) and 5 (d)). *Figure 5*(a) shows a path planning which avoids obstacles such as *buildings*, *walls*, but does not take into account the terrain characteristics. Therefore, this path crosses areas coloured in red which represent steep slopes. However, *Figure 5*(b) respects both the terrain and the obstacles in the IVGE. To illustrate path planning towards a target area qualified by one or several semantics, *Figure 5*(c) shows the computed path to reach the marina (the marina is coloured in blue at the top of the figure). This path avoids obstacles such as *buildings*, *walls*, but does not take into account the terrain characteristics. *Figure 5*(d) avoids steep slopes (coloured in red) as well as obstacles situated in the IVGE and reaches a place identified by the semantic information (marina). Finally, in order to highlight the outcomes of the path optimisation process, we randomly selected 19 starting and destination positions in the IVGE. For each pair of positions, we compared the original computed path length with the optimised path length. *Figure 7* depicts the comparison of the non optimised computed path length and the optimised

path length. It shows how the optimisation process reduces the computed path length by an average of 16%.

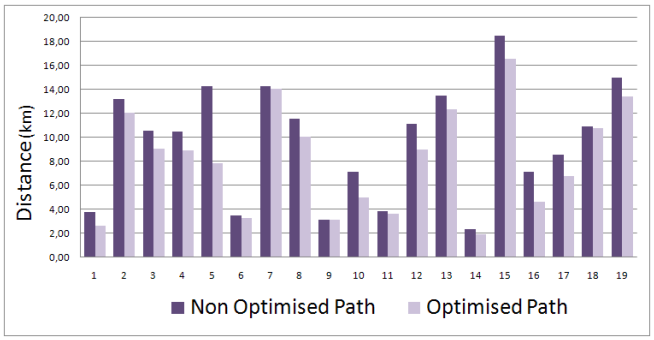



*Figure 7:* Optimised versus non-optimised paths lengths.

## VI. Navigation

The geometrically precise spatial subdivision along with the topologic abstraction of the virtual geographic environment are not sufficient to handle the navigation of several moving agents populating the same IVGE. A structure and a mechanism allowing for dynamic collisions detection and avoidance is necessary to achieve consistent motion planning of moving agents in IVGE. In this section, we first introduce the concept of neighborhood graph (NG) for the support of agent navigation. Next, we detail the algorithm that we propose to build an NG while taking into account obstacles such as walls and fences as well as terrain elevation.

*Neighborhood Graph:* A *neighborhood graph* (NG) consists of a data structure reflecting the relative positions of moving agents while taking into account obstacles located in the IVGE. Two entities are considered neighbors if they are not separated by any obstacle. Since the NG is based on moving agents' positions, it should be updated as rapidly as the movement of agents. Therefore, its computation complexity must be optimised. Moreover, if we make no assumption about the perception distance and the angle of the agent's field of view, the complexity of building the NG should only depend on the number of agents rather than on their relative distances. Based on these assumptions, we define our NG using a *3D Delaunay Triangulation* (3D-DT) of moving agents (*dynamic information*) filtered using both the description of static obstacles situated in the IVGE such as walls and obstructions resulting for the terrain's elevation (*static information*). Figure 8 shows an example of construction of an NG considering obstacles and obstructions within the IVGE. In the following sud-section, we propose an algorithm to build NGs and we analyse its complexity.

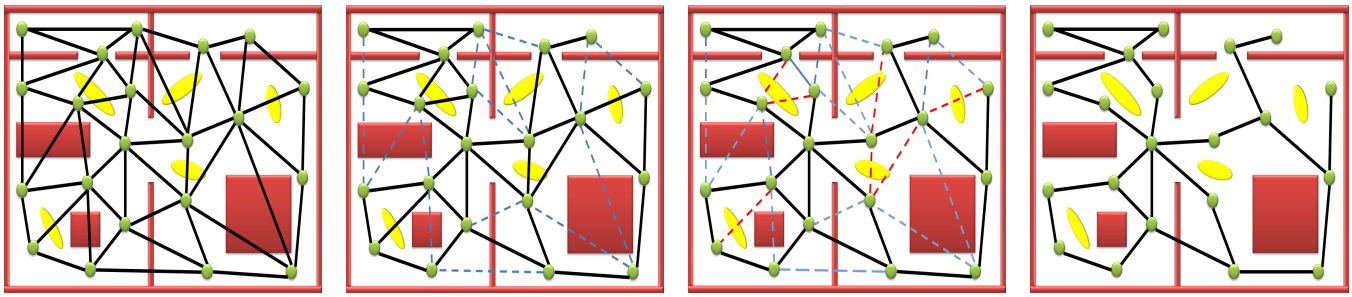
*Algorithm:* In this section, we present a step-by-step description of our algorithm to build NGs. In a first step, the spatial positions of the moving agents are collected and constitute the set of points to triangulate. Let $n$ points be given by their Cartesian coordinates $p_1(x_1, y_1, z_1)$, $p_2(x_2, y_2, z_2)$, ..., $p_n(x_n, y_n, z_n)$. The algorithm is based on three steps. Step 1: Compute the 3D-DT; Step 2: For each edge $E_{i,j}$ of the DT linking a pair of points ($P_i$, $P_j$), verify if this edge crosses an area defined as an obstacle in the IVGE. If yes, remove $E_{i,j}$. Step 3: For each edge $E_{i,j}$ of the DT linking a pair of points ($P_i$, $P_j$), verify if $P_i$ and $P_j$ are obstructed as a result of the terrain's elevation. If yes, remove $E_{i,j}$.

The complexity of construction of the 3D-DT for $n$ moving agents is of the order of $O(n \ln n)$, making it usable for a large number of moving agents (Figure 8(a)). In the second step, for each edge $E_{i,j}$ of the 3D-DT, a verification is computed to ensure the visibility (free of environment obstacles) between the moving agents (Figure 8(b)). In the third step, for each edge $E_{i,j}$ of the 3D-DT, a verification is computed to ensure the visibility (free of environment obstructions resulting from terrain elevation) between the moving agents (Figure 8(c)). If $E_{i,j}$ is not free of obstacles and obstructions, the edge is deleted from the 3D-DT (Figure 8(d)). Let us now analyse the complexity of our algorithm. The 3D-DT (Step 1) can be computed in $O(n\log n)$ running time [2]. In Step 2, for each edge $E_{i,j}$ considered, $O(n)$ verifications are computed to ensure that $E_{i,j}$ is free of environment obstacles. These verifications have a linear complexity which only depends on the number of vertices (corresponding to the moving agents). Step 2 runs in $O(n)$ time. In Step 3, for each edge $E_{i,j}$ considered, $O(n)$ verifications are computed to ensure that $E_{i,j}$ is free of obstructions due to the terrain elevation. Step 3 also runs in $O(n)$ time. Since Step 1 is $O(n\log n)$, the complexity of our NG algorithm is dominated by Step 1 and thus of $O(n\log n)$.

## VII. Discussion

In order to reduce the search space, we proposed an HTG that groups convex cells and groups of cells. Hence, each abstract node at level $i$ contains a subset of this graph at level $i-1$, composed by at least one node or abstract node. The extraction of this HTG only requires an acceptable one-time computation cost and a low memory overhead. Despite the reduction of the number of nodes, this technique raises two application-dependent issues that must be addressed: ***hierarchical traversal cost*** and ***information richness***.

First, the *hierarchical traversal cost* increases with each grouping, which might limit the performance of the search space reduction brought by the hierarchical representation. Indeed, despite the number of levels of the HTG, the path planning process provides moving agents with a set of convex cells (belonging to level 0) to pass through in order to reach the final destination. This means that the path planning process must inevitably traverse the HTG from top to bottom



(a) (b) (c) (d)

*Figure 8:* Generation of the neighborhood graph (NG): (a) initial 3D-DT using the moving agents positions (green circles); (b) filter of 3D-DT considering obstacles (red shapes) in the IVGE; (c) filter of 3D-DT considering terrain elevation (yellow shapes) in the IVGE.

in order to compute such a set of cells.

Second, the *information richness* decreases with each grouping level, which could lead to useless additional abstraction levels that may not improve the decision-making of the HPP algorithm. Indeed, the more potential sub-paths an abstract node contains, the less its choice influences the path planning process. Therefore, the determination of the number of topologic abstraction levels must be carefully analyzed with respect to these two critical issues in addition to the application requirements.

Another important aspect of our IVGE is its capability to represent geographic environments which are distributed in space. By using the HTG, our model is capable of representing portions of geographic environments which are not adjacent in space. For example, consider the problem of traveling by car from Quebec city (QC, Canada) to New York (NY, USA). We need to compute the shortest (minimum distance) path from a given address in Quebec city, such as 312 *Marie-Louise*, to a given address in New York city, let us say 1213 4*th* Avenue, *Brooklyn*. Given a detailed description of the geographic environment showing all roads annotated with driving distances, a classic planner can compute such a travel route. However, this might be an expensive computation, given the large size of the description of the geographic environment. This problem may be solved in a three-step process. First, we compute the path from 312 *Marie-Louise* to a major highway leading out of Quebec city. Second, we compute the path from Quebec to the boundaries of New York. Third, we compute the path from the incoming highway to 1213 4*th* Avenue, *Brooklyn*. Assuming that the second path is mostly composed of highways and can be quantified (distance and travel time), it is easy to model this path using a *conceptual node* in our hierarchical topologic graph. A conceptual node allows for linking spatially distributed geographic environments and hence allows us to accurately compute optimal paths with respect to these environment characteristics.

In contrast to Lamarche and Donikian's NG [12] which only takes into account static obstacles such as walls and fences (Figure 8(b)), our NG model also includes obstructions resulting from the terrain's elevation (Figure 8(c)). Lamarche and Donikian's NG is based on a 2D-DT implementing the algorithm proposed in [4] whose computation cost is of $O(n\log n)$ [12]. However, Attali *et al.* proposed an optimised algorithm to build a 3D-DT which also runs in $O(n\log n)$ [2]. Our NG extends Lamarche and Donikian's approach with respect to the algorithm's complexity ($O(n\log n)$). Our NG takes into account the terrain's elevation and uses Attali's optimised 3D-DT and thus runs in $O(n\log n)$. In addition to its good properties in terms of computation complexity, the DT also has good topological properties. Indeed, it ensures that each point is connected to its nearest neighbor. An NG inherits from this property by adding the concept of filtering visibility. We define the $k-direct$ neighborhood $V_k(E)$ of an agent $E$ as all agents related to $E$ by $k$ arcs in the NG. This set contains the nearest visible neighbors to an agent within $k$ hops from $E$, considering the NG. This property shows that for the collision detection purposes, only agents belonging to $V_1(E)$ (also called *immediate neighbors*) need to be tested. On the other hand, according to the method of construction, the $k-direct$ neighborhood adapts to the density of population. For example, in a dense environment, the $k-direct$ neighborhood contains a set of agents which are visible and close to the agent $E$, and in environment of low density, it contains a set of remote visible agents. The $k$ parameter allows to specify the number of hops while accessing the agent neighbors by agent type.

## VIII. Conclusion and Future Works

In this paper, we proposed an accurate and automated approach for the generation of semantically enhanced and geometrically precise virtual geographic environments using GIS data. This novel approach offers several advantages. First, the description of the IVGE is realistic since it is based

on standard GIS data and accurate because it is produced by an exact spatial decomposition technique which uses data in a vector format. Hence, this description preserves both the geometric and the topological characteristics of the geographic environment and enables a graph-based description of the virtual environment enhanced with semantics. The main outcome of such a semantically enhanced and geometrically precise virtual geographic environment concerns agents' situated reasoning capabilities such as path planning and navigation in large-scale and complex geographic environments. We proposed a hierarchical path planning algorithm (using *Dijkstra* and *A**) which takes advantage of our IVGE model to provide paths which take into account the agents' and environment's characteristics. We also proposed an algorithm to build neighborhood graphs, a graph-based structure used by moving agents to support navigation (collision detection and avoidance).

We are currently working on further improvements of the IVGE description by integrating enriched knowledge representations (called *the environment knowledge*) using *Conceptual Graphs* aimed at assisting situated agents' interactions with the IVGE and helping them achieve their goals. The goal of the environment knowledge integration is to extend the agents' knowledge about their surrounding environment. We are also working on the extension of the neighborhood graph concept to support agents' perception capabilities within the IVGE. The above-mentioned contributions of our model offer new opportunities for many applications in a variety of application domains including the entertainment industry (games and movies), security planning and crowd management (planning events involving large crowds), and environment monitoring in natural environments.

## ACKNOWLEDGEMENT

Mehdi Mekni benefited from a PDF scholarship granted by FQRNT (***F**onds **Q**uébécois de la **R**echerche sur la **N**ature et les **T**echnologies*).

## REFERENCES

[1] W. Ali and B. Moulin. 2D-3D multiagent geosimulation with knowledge-based agents of customers' shopping behavior in a shopping mall. In *Spatial Information Theory*, pages 445–458. Elsevier, 2005.

[2] D. Attali, J.-D. Boissonnat, and A. Lieutier. Complexity of the Delaunay triangulation of points on surfaces: the smooth case. In *SCG '03: Proceedings of the Nineteenth Annual Symposium on Computational Geometry*, pages 201–210, New York, NY, USA, 2003. ACM.

[3] I. Benenson and P. Torrens. *Geosimulation: Automata-Based Modeling of Urban Phenomena*. John Wiley and Sons Inc., 2004.

[4] J.-D. Boissonnat and M. Yvinec. *Algorithmic Geometry*. Cambridge University Press, New York, NY, USA, 1998.

[5] A. Botea, M. Müller, and J. Schaeffer. Near optimal hierarchical path-finding. *Journal of Game Development*, 1:7–28, 2004.

[6] D. Demyen and M. Buro. Efficient triangulation-based pathfinding. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference (AAAI'06)*, Boston, Massachusetts, USA, July 16-20 2006.

[7] S. Donikian and S. Paris. Towards embodied and situated virtual humans. In *Motion in Games*, pages 51–62, 2008.

[8] N. Farenc, R. Boulic, and D. Thalmann. An informed environment dedicated to the simulation of virtual humans in urban context. In P. Brunet and R. Scopigno, editors, *Computer Graphics Forum (Eurographics '99)*, volume 18(3), pages 309–318. The Eurographics Association and Blackwell Publishers, 1999.

[9] H. Haddad, M. Khatib, S. Lacroix, and R. Chatila. Reactive navigation in outdoor environments using potential fields. In *Proceedings of the International Conference on Robotics and Automation*, pages 1232–1237, Leuven, Belgium, May 1998. IEEE.

[10] D. Harabor and A. Botea. Hierarchical path planning for multi-size agents in heterogenous environments. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS'08)*, Sydney, Australia, September 14-18 2008.

[11] D. Helbing, I. Farkas, and T. Vicsek. Simulating dynamical features of escape panic. *Nature*, 407:487, 2000.

[12] F. Lamarche and S. Donikian. Crowds of virtual humans: a new approach for real time navigation in complex and structured environments. *Computer Graphics Forum, Eurographics'04*, 2004.

[13] S. LaValle. *Planning Algorithms*. Cambridge University Press., Cambridge, 2006.

[14] J. Lengyel, M. Reichert, B. R. Donald, and D. P. Greenberg. Real-time robot motion planning using rasterizing computer graphics hardware. In *SIGGRAPH '90: Proceedings of the 17th Annual Conference on Computer Graphics and Interactive Techniques*, pages 327–335, New York, NY, USA, 1990. ACM.

[15] J.-E. Marvie, J. Perret, and K. Bouatouch. Remote interactive walkthrough of city models. In *Proceedings of the 11th Pacific Conference on Computer Graphics and Applications (PG'03)*, pages 389–393, Oct. 2003.

[16] N. Nilsson. *Principles of Artificial Intelligence*. Springer-Verlag, Berlin ; Heidelberg ; New York, third edition, 1982.

[17] S. Paris. *Characterisation of the levels of services and modeling of the movement of people inside exchange areas*. PhD thesis, Université de Rennes 1, October 2007.

[18] S. Quinlan and O. Khatib. Elastic bands: Connecting path planning and control. In *ICRA (2)*, pages 802–807, 1993.

[19] S. Rodriguez, V. Hilaire, S. Galland, and A. Koukam. Holonic modeling of environments for situated multi-agent systems. In *Environments for Multi-Agent Systems II*, pages 18–31. 2006.

[20] W. Shao and D. Terzopoulos. Environmental modeling for autonomous virtual pedestrians. *Digital Human Modeling for Design and Engineering Symposium*, 2005.

[21] A. Sud, R. Gayle, E. Andersen, S. Guy, M. Lin, and D. Manocha. Real-time navigation of independent agents using adaptive roadmaps. In *SIGGRAPH '08: ACM SIGGRAPH 2008 classes*, pages 1–10, New York, NY, USA, 2008. ACM.

[22] G. Thomas and S. Donikian. Virtual humans animation in informed urban environments. *Computer Animation 2000*, pages 112–119, 2000.

[23] M. Wooldridge. *Introduction to Multiagent Systems*. John Wiley and Sons Inc., London, UK, 2001.

[24] Y. Yang and O. Brock. Elastic roadmaps: Globally task-consistent motion for autonomous mobile manipulation in dynamic environments. In *Robotics: Science and Systems*, 2006.

[25] J. Zhu, J. Gong, H. Lin, W. Li, J. Zhang, and X. Wu. Spatial analysis services in virtual geographic environment based on grid technologies. *MIPPR 2005: Geospatial Information, Data Mining, and Applications*, 6045(1):604–615, 2005.

# Enhanced User Interaction to Qualify Web Resources by the Example of Tag Rating in Folksonomies

Monika Steinberg, Orhan Sarioglu, Jürgen Brehm
Institute of Systems Engineering - System and Computer Architecture
Hannover, Germany
[steinberg, brehm]@sra.uni-hannover.de, orhansarioglu@freenet.de

*Abstract* - **The Web offers autonomous and frequently useful resources in growing manner. User Generated Content (UGC) like Wikis, Weblogs or Webfeeds often do not have one responsible authorship or declared experts who checked the created content for e.g., accuracy, availability, objectivity or reputation. The user is not able easily, to control the quality of the content he receives. If we want to utilize the distributed information flood as a linked knowledge base for higher-layered applications – e.g., for knowledge transfer and learning – information quality (iq) is a very important and complex aspect to analyze, personalize and annotate resources [1]. In general, low information quality is one of the main discriminators of data sources on the Web [2]. Assessing information quality with measurable terms can offer a personalized and smart view on a broad, global knowledge base. We developed the qKAI application framework [3] to utilize available, distributed data sets in a practically manner. In the following, we present our adaption of information quality aspects to qualify Web resources based on a three-level assessment model. We deploy knowledge-related iq-criteria as tool to implement iq-mechanisms stepwise into the qKAI framework. Here, we exemplify selected criteria of information quality in qKAI like relevance or accuracy. We derived assessment methods for certain iq-criteria enabling rich, game-based user interaction and semantic resource annotation. Open Content is embedded into knowledge games to increase the users' access and learning motivation. As side effect the resources' quality is enhanced stepwise by ongoing user interaction. By the example of image tag rating in folksonomies we demonstrate a practicable use case for qualifying web resources by keyword-oriented group search and game-based tag ranking in detail.**

***Keywords - Information Quality, Folksonomy, Open Content, Semantic Annotation, Knowledge Transfer.***

## I. INTRODUCTION

If we want to embed Web content into knowledge transfer and learning, the question about the data's' quality is indispensable. Information quality (iq) is an important concern if we want to build knowledge out of information towards education. Currently, Web users are claiming for more sophisticated content and less triviality [4]. To utilize autonomous web resources qualitative assessment of the broad information load becomes more and more important.

To let users interact with Open Content [5] out of distributed web resources, enhanced inquiry, selection, storage and buffering are important prerequisites. Nevertheless, statements of the resources iq enhance its fitness for use. The more we know about a resource, the better we can reuse it. We developed the qKAI application framework (qualifying Knowledge Acquisition and Inquiry) [3] - a service-oriented, generic and hybrid approach combining knowledge related offers for convenient reuse. As part of the qKAI application framework, we implemented the qKAI hybrid data layer to acquire, store and represent Open Content out of distributed resources. In qKAI Open Content is boosted as an inherent part of higher-layered applications in knowledge and information transfer via standard tasks of knowledge engineering and augmented user interaction. Especially regarding smart user interaction, we have to offer user interfaces with high scores in certain information quality criteria. If we get to know about a resource that it contains Chinese text by analyzing its metadata, we can deduce that its "*understandability*" is almost not ideal for European users. There are lots of small hints and tasks that are very helpful altogether to assess and enhance information quality aspects of Open Content. In the following, we introduce the meaning of information quality in qKAI exemplified with selected criteria of iq. We explain the relation between these criteria, qKAI data and interaction issues with Open Content.

### A. Structure of this contribution

First, we introduce some further background. In Section 2 follows the state of the art. Section 3 gives an overview of information quality (iq) criteria. Section 4 offers some more details about assessing Web contents' quality. Section 5 shows how to combine traditional information quality metrics with enhanced user interaction, Section 6 exemplifies the use case of pictures' relevance in folksonomies, Section 7 gives a short analyzes of tagging systems. Section 8 explains our derived approach: keyword-oriented group search for tag ranking in folksonomies. Section 9 shows our game-based tag rating approach qRANK. Section 10 illustrates some evaluation results regarding our approaches versus the standard Flickr keyword search. Section 11 offers further application scenarios and use cases. At least this contribution ends up with conclusion and outlook in Section 12.

### B. Utilizing Open Content for knowledge transfer and learning

The qKAI application framework serves as basis to develop rich user interaction with Open Content [6] upon it. Actually, we implement and evaluate knowledge visualization and game prototypes. qMAP acts as an

interface to visualize and interact with Open Content like images, texts or videos geographically on a map. qMATCH offers the user term-image or term-term assignment questions out of Flickr content. qCHUNK lets the user guess Wikipedia articles while he gets presented chunks out of them. qMAP is a geocoded, map-based gaming board to visualize Open Content like Wikipedia articles or Flickr images. Some examples are shortly presented in Chapter 6. We see game-based interaction as a use-case with high design and interaction receivables that is well suited to evaluate enhanced interaction with Open Content exemplary.

### C. Assessing the information quality of autonomous web resources

"*Information quality (iq) is one of the main discriminators of data and data sources on the Web. ... The autonomy of Web data sources renders it necessary and useful to consider their quality when accessing them and integrating their data.*" [2].

Information quality is often described as "*fitness for use*" [7] in the relevant literature. Metadata plays an important role for the determination of iq-criteria. Information quality is to a great extend subjective, because we have to mention multi-dimensional criteria while assessing context-, user- and task-dependent. Subjective dimensions of iq must be assessed by the help of user interaction [2]. User interaction can be basic, direct or indirect feedback.

… "*Many iq-criteria are of subjective nature and can therefore not be assessed automatically, i.e., independently and without help of the user.*"… [2]

Because iq is often subjective, task- and context-dependent, **user interaction** plays a very important role while assessing subjective iq-criteria. To let users rate and rank content according to certain iq-criteria, questionnaires are widely used.

### D. Semantic annotation of resources

qKAI delivers an URI about every resource it utilizes in RDF [8] representation. Semantic interlinking between the provenance resource and the new, annotating qKAI URI connects the URIs following Linked Data paradigms. Semantic interlinking allows following all references (links) automatically. HTML for example does not offer this ability.

### E. A global interaction rewarding model (GIAR)

An ontology-based interaction rewarding model (GIAR) is work in progress. qKAI rewards any kind of interaction with resources and other users to increase user participation and incentive. Therefore, we are designing a catalogue of interaction tasks and order them according to domain, type and further iq- criteria. For every interaction the user earns points according to a global point and level system like in game-based scenarios. We are rewarding external activity also, like e.g., listening to music at Last.fm, making friends at Facebook or tweets at Twitter. Every interaction is stored in a personal profile file that builds knowledge-related reputation step by step. Every resource has its own transaction and interaction protocol (see Figure 3 in Chapter 6). The protocol can be statistically evaluated to enable automated ranking, rating and deriving further iq-criteria. A social interaction rewarding community is under development to visualize the global interaction rewarding concept.

## II. State of the Art and related Work

Wang [9], Naumann [2] and Bizer [13] a.o. offer comprehensive research work about categorization, definition of information quality and related vocabulary in the domain of webbased information system. Wikipedia [10] has its own quality assessment deploying a review mode by authors. Freebase [11] allows the user to rearrange, connect, correct or annotate available resources. Rating, ranking and recommendation at Amazon [12] are good examples for enhanced user interaction to qualify content. Flickr offers properties related to a picture that enable to rate a photos quality. Tagging allows users to restructure and weight their knowledge in a self-controlled way. Revyu [14] allows the users to rank and rate everything. In qKAI we will integrate Revyu by querying whether a resource is annotated by Revyu yet. The reputation of a thing, person or resource in qKAI is increased if there is a Revyu entry about it. The existence of available interlinked context information in e.g., other web platforms is a first and simple step to determine information quality of resources according to scores.

## III. Information Quality Criteria and Open Web Content

The "*fitness for use*" can depend on numerous factors like actuality, believability, completeness or relevance. Not all single criteria are assessable independent from each other [13]. Next to several further properties the most important criteria of information quality in web applications are actuality, reputation, believability and accuracy of content.

In contrast to processes inside of enclosed organizations that analyze iq as cyclic management task the assessment of iq in the Web relies on autonomous information providers in an open information space. Therefore, in webbased systems IQ is assessed by the help of user interaction to determine the "*fitness for use*" of an information source for the specific task on hand [13]. Social aspects of iq especially in the context of Web 2.0 are reputation and trust of the author.

Important for the believability of information is the reputation of the creator. Every user has his own opinion based upon own experience or the experience in his knowledge circle. All experiences that are made with resources in qKAI are logged in history protocols. Different opinions about the reliability or trustworthiness of single actors regarding certain themes emerge. Personalized knowledge views can be deduced this way.

There are trust metrics and policies for **reputation-based systems** available in literature and research [15] that can be implemented next to **interaction-based** and **metadata-relying** metrics.

### A. Categorizing Information Quality

The categorization of information quality is in respective literature available according to various criteria and dimensions [16]. We did not find much about generic interaction components to assess ongoing iq in web-based knowledge systems by online assessment [17] components with game-based features. We see especially the combination of reputation-based and global metrics as promising first step towards an incentive and motivating way to assess iq sustainable.

TABLE I. IQ CITERIA AND THEIR CLASSIFICATION FOR AUTONOMOUS INFORMATION SYSTEMS BASED ON C. BIZER'S CATEGORIZATION [13]

| Category | Criteria/Dimension | Objective/ subjective |
|---|---|---|
| **Intrinsic criteria** (Independent of the user's context) | Accuracy* | objective |
| | Consistency | objective |
| | Objectivity | objective |
| | Timeliness | objective |
| **Contextual criteria** (Context, task and user dependent) | Believability | subjective |
| | Completeness | subjective |
| | Understandability | subjective |
| | Relevancy | subjective |
| | Reputation | subjective |
| | Verifiability | subjective |
| | Amount of Data | subjective |
| **Representational criteria** | Interpretability | subjective |
| | Rep. Conciseness | subjective |
| | Rep. Consistency | objective |
| **Accessibility criteria** | Availability | objective |
| | Response Time | objective |
| | Security | objective |

*Accuracy is interpreted in a bias way in qKAI: On the one side, we have to assess the data accuracy, on the other side we speak of semantically and syntactically correct information. The last one can only be assessed by enhanced user interaction of experts or collective intelligence approaches (Wisdom of crowds).

Accuracy is defined as the percentage of data without data errors, such as non unique keys or out of range values. Mohan et al. give a list of possible data errors [2].

### B. Iq-criteria for the qKAI system domain

It is not practicable to measure all available iq-criteria at once. We have to select the most important criteria for our domain. In qKAI we have a strong focus on knowledge transfer with smart interaction. To offer knowledge-related content, we have to fulfill e.g., semantically correctness of factual data. We interpret semantically correctness as one aspect of accuracy. Accuracy is defined as the degree of correctness and precision with which information in an information system represents states of the real world [14].

**Figure 1** shows the actually most important iq-criteria in the qKAI system domain.

Technical or also called accessibility criteria like availability, response time or security depend almost on soft- and hardware concerns. We developed the qKAI hybrid data layer as part of the qKAI application framework to offer good results for these technically oriented criteria on an affordable Quadcore-platform. qKAI is suitable to search and explore distributed resources in an effective manner and represents our ongoing and enhanced research toward hybrid data management for distributed resources with rich interaction on top of it. To reach good results in the frontend, the backend – including the data layer - has to be suitable for this purpose. E.g., if a user waits too long, to get first search

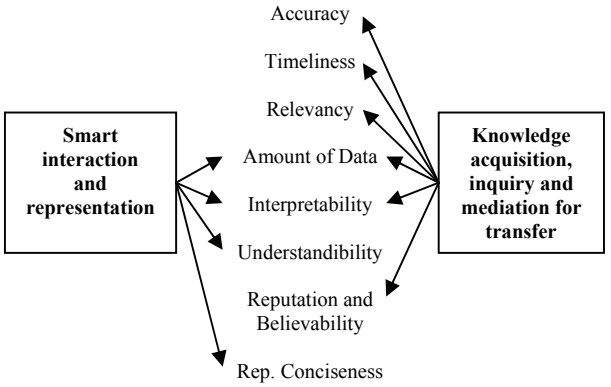



Figure 1. Most relevant iq-criteria for the qKAI system domain: knowledge transfer and smart interaction based on autonomous resources

results, the motivation to ongoing interaction will rapidly increase. The iq-criteria "*response time*" and "*availability*" have to be enhanced by technically aspects like hard- or software requirements.

### C. Reputation as quality criteria and for users' motivation

Reputation can be seen as the sum of single experiences and expectation about trustworthiness and competence of a person, a group or an organization. Reputation has much to do with image and status of a person or a thing and is an important factor in online communities, where trust and reliability come into play. Most online communities that collect feedback to qualify content do not offer any incentive to rate and rank. The missing motivation of users to interact on the content is an essential problem, because there are no rational reasons to participate sustainably and the chance is taken to let other users do the ratings [18]. Creating and enhancing the own reputation is next to the simple fun [19] a good motivator to embed online users into to content-related participation without material incentive [18] [13]. Ebay and Amazon are successful examples for building reputation by users' feedback. In qKAI, the reputation of users is stored implicitly in their personal profile and increases with every kind of interaction on Open Content. A resources reputation is stored in their semantically linked qKAI annotation URI and is also increased by any interaction or analyzes, the resource is involved.

## IV. ASSESSING THE QUALITY OF WEB CONTENT

…"*Information quality assessment is the process of assigning numerical values (iq-scores) to iq-criteria. An iq-score reflects one aspect of information quality of a set of data items."* …[1]

To assess the iq of information sources, a scoring function calculates assessment scores form the collected ratings. The scoring function decides which ratings are taken into account and might assign different weights to ratings. Which criteria to take for a specific rating should be adjustable by the user and his task on hand? Our research showed the following classifications and assessment models to be most suitable for qKAI and webbased information systems with knowledge-related concerns in general. Naumann identified three main factors the quality of information is influenced by in his query-oriented approach:

- the perception of the user (the subject of a query),
- the data itself (the object of a query),
- the process of accessing the data (the predicate of a query) [2].

C. Bizer [13] derived three levels of information quality metrics in web-based information systems:

- **Content-Based Metrics** use information to be assessed itself as quality indicator. The methods analyze information itself or compare information with related information.
- **Context-Based Metrics** employ meta-information about the information content and the circumstances in which information was created, e.g., who said what and when, as quality indicator.
- **Rating-Based Metrics** rely on explicit ratings about information itself, information sources, or information providers. Ratings may originate from the information consumer herself, other information consumers, or domain experts.

We adjusted these three levels to assess iq for qKAI needs to **first, second and third level assessment** divided into **Metadata analysis, user interaction** and **intelligent analysis**. There is no absolute quality, but we can compare resources with each other (Open World Assumption) and weight them based on the amount and structure of metadata, for example. Enrichment of a resource happens in a corresponding qKAI URI by semantic interlinking and annotation. Ranking according to available metadata properties or interaction history is possible too.

### A. First level assessment: Metadata analysis

According to Bizer this level enables **Context-based assessment** of metadata directly related to a resource like format, timeliness, author, provenance or language, which can be automatically detected. Metadata can be seen as a quality feature. The more metadata we are extracting, the better we get to know the content. In qKAI we are implementing the support of Aperture [20] to fetch e.g., Dublin core elements [21] like listed in **Table 2**.

TABLE II. EXEMPLARY DUBLIN CORE ELEMT SET FOR METADATA [21]

| Element | Definition and recommended value formats |
|---|---|
| Title | A name given to the resource.<br>Value format: Free text. |
| Creator | An entity primarily responsible for creating the content of the resource. Value format: Name as free text. |
| Subject | A topic of the content of the resource.<br>Value formats: Library of Congress Subject Headings (LCSH), Medical Subject Headings (MeSH), Dewy Decimal Classification (DDC). |
| Description | An account of the content of the resource. Value format: Free text. |
| Publisher | An entity responsible for making the resource available. Value format: Name as free text. |
| Contributor | An entity responsible for making contributions to the content of the resource. Value format: Name as free text. |
| Date | The date when the resource was created or made available. Value Format: W3C-DTF. |
| Type | The nature or genre of the content of the resource.<br>Value Format: DCMI Type Vocabulary. |
| Format | The physical or digital manifestation of the resource.<br>Value Format: MIME-Type. |
| … | … |

Comparable iq scores can be derived out of adjustable quality policies like e.g., available metadata property count: The less metadata properties a resource contains, the smaller is its iq score in believability or reputation. Even provenance and timeliness are very important aspects concerning trust in a resources' content. Information about the author is also very relevant for the resources quality. A user with high personal scores in certain knowledge domains has high reputation in this area. We can speak of local reputation here, because it is dependent the same way, the iq-criteria are, from task, user and context.

### B. Second level assessment: User interaction

We allocate criteria here that can be assessed with the help of user interaction. Questionnaires are often used to get feedback from the user for this purpose. According to Bizer this is called **Rating-based assessment.**

The user can help e.g., to enhance accuracy even regarding semantically correctness. To evaluate factual knowledge like "*Berlin lies at the Spree*" or "*Hanover is the capital of Lower Saxony*", we see user rating and ranking following the established Web 2.0 manner as an effective solution to mark wrong content and to rank valuable or popular content step by step. Next to this crowd sourcing community approach we offer role- and level-based quality control mechanisms. Lecturers earn rewards while rating and creating educational resources out of Open Content; students earn rewards while answering questions, managing gaming tasks, exploring further content or ranking their favorites. Step-wise content can be qualified this way. Resources are marked following their quality level as **reviewed, proofed**

**or not yet qualified** to enable embedding in different levels of knowledge transfer and learning. Integrating online assessment components like multiple-choice or assignment question types into social oriented software seems to be a new approach – as far as we know. Although, online assessment and rating mechanisms have many things in common and can be complementary, their combination is not mentioned so far.

### C. Third level assessment: Intelligent analysis

By **Content-based assessment** employing Natural Language Processing to detect some more information hidden inside a resource. Aperture [20] and Virtuoso Spongers [22], for example, enable comprehensive solutions for these tasks. In case if more text engineering is needed, there are comprehensive solutions for standard Natural Language Processing (NLP) tasks (e.g., by OpenNLP [23]) to perform sentence detection, NER (Named Entity Recognition), POS (Part-Of-Speech) tagging or even semantic chunking. **Table 1.** shows the related iq-criteria from relevant literature. If we talk about information quality, we also talk about user preferences and personalization. It is obvious that many of the iq-criteria are relevant while user interaction takes place, because they are subjective – user, task and context dependent. Most of the iq-criteria have direct impact on the users' interaction. There are only a few iq-criteria like "*amount of data*" or "*completeness*" that can be assessed with little or no user interaction at all. Even technical criteria influence usability, ease of use and user motivation elementary. Without fulfilling e.g., technical criteria in a sufficient way, smart interaction is not possible at the user side. Altogether, the 2nd level of our qualifying model with strong focus on user interaction is the most important and influential one if we want to determine relevant, but subjective iq scores.

## V. IQ ASSESSMENT WITH THE HELP OF ENHANCED USER INTERACTION

Incentive for user participation is implemented as globally rewarding system of any interaction in qKAI (qPOINT, qRANK). **Table 3** shows interaction types, their assigned reward in form of gaming points and improvable iq-criteria. Every interaction is based on a resource. We are implementing different types of interaction like described in the following.

TABLE III. INTERACTION TASKS, ASSIGNED REWARDING POINTS AND IMPROVABLE IQ-CRITERIA

| Interaction | Reward | Improvable iq-criteria |
|---|---|---|
| **Edit** | +50 points | Accuracy, consistency, objectivity, timeliness, believability, reputation, completeness, understandability |
| **Create** | +100 points | Completeness, accuracy, verifiability, amount of data |
| **Annotate/ add/interlink** | +50 points | Completeness, accuracy, verifiability, amount of data, interpretability, understandability |
| **Rate/rank** | +10 points | Relevancy, accuracy, believability, reputation, objectivity, interpretability, understandability, rep. conciseness |

### A. Simple and direct feedback

Like in common surveys and evaluation, rating happens by questionnaires with predefined scores. These ratings can evaluate persons, resources or knowledge units.

### B. Enhanced feedback and game-based interaction

Every resource that is visualized or just queried by qKAI can be rated and ranked by user interaction or automated metrics like metadata detection. The more a resource is requested, the more statitistically data we gain. The more we know about a resource, the better we can personalize its usage.

Next to edit, create, annotate, add, interlink and rate resources and users we offer the following game-based options. qKAI jokers allow game-based functionality to add additional sources and to qualify metadata by rating and ranking input to the qKAI knowledge base. Playing the "*Know-it-all-Joker*" bounds the user to add a source (or information) that proves contrary statements. The "*Nonsense-Joker*" marks an information unit as semantically wrong or inconsistent and defers it to review mode by other qKAI users. The "*Hint-Joker*" allows looking up related sources or other users' answers as solution suggestion. The "*Explorer-Joker*" allows exploring the right answer on the web outside of qKAI during a predefined time. The "*History-Joker*" enables lookups in played answers, ratings of other users by logged interaction and transaction protocols. Statistical protocol analysis is suitable to infer further metadata.

### C. Indirect and automated feedback

History protocols and interaction recording allows to deduce statistically results for rating and ranking purpose. Therefore, Simple Scoring Functions, Collaborative Filtering, Web-of Trust algorithms or Flow Models can be deployed in the future.

## VI. THE RELEVANCE OF PICTURES IN FOLKSONOMIES

In this section we introduce one of our example use cases to enhance and determine the quality of Open Content by **collective intelligence**. Tagging is very popular in online communities these days. Everybody can participate in tagging content. Tags offer a wide range of keywords but are subjective as well and might be confusing sometimes.

### A. Relevance of pictures

Focus here is the quality of the images found on the web. With Flickr [40] a highly demanding data source with more than two billion images and over two million new images per day is given. A crucial problem that has emerged during the study was the relevance of the found images. Many images that are found with the help of the Flickr web service do not clearly correspond to the search term. They do not deliver the desired content. The challenge that arises from this is the

automatic sorting of images according to their relevance. Flickr offers a very comprehensive interface (API) which allows more possibilities than the pure web service. We developed a small application called Flickr-analyzer that is used for analytical purposes. The search for clean images, in contrast to text or text-picture combinations is difficult because they contain too little information to be found. [24]

"*There are many resources which are not searchable in folksonomies because they do not contain most of the relevant tags*" [25].

One way to facilitate the search of images on the Web is additional metadata e.g., by adding tags of our own choice. This kind of annotation is different from professional annotations in that they do not use notations and relations. Basically annotations facilitate the search and navigation of resources. The common form of this annotation is referred to in the latest generation of the Web as **collaborative tagging**. Services that allow this type of metadata generation are known as **tagging systems**. The most famous among them are Flickr [40], YouTube [44] and Del.icio.us [45].

### *B. Tagging and tagging systems*

If user index resources with additional keywords called tags, this is called "*tagging*". There are two types of tags: normal tags and machine tags. The former are from users randomly selected keywords that reflect mostly the image content or additional information about the resources. Machine tags are machine-generated tags. These include auto-tagging and tags, which have a certain shape. Geo-tags are information indicating the geographical coordinates of the origin of the pictures or the coordinates of objects, which are shown in the pictures. Web 2.0 services that allow collaborative tagging are known as tagging systems [26]. Tagging not only organizes the resources in tagging systems in a better way, but also means that a collaborative network is formed.

"*Social tagging is used by users to build both its own network, as well as the network to" watch ", and get as new sources for the topic areas of interest*" [27]

### *C. Geo-Tagging*

Geo-tagging is composed of two words, "*Geo*" and "*tagging*" and describes the geographic positioning of e.g., images. Many images are from a GPS receiver located at the camera automatically. This means images will be automatically marked with longitude and latitude. In Flickr, users geo-tagged their photos on a specific format: geo: lon = 13.127787 geo: lat = 52.393684. This allows an image to be found with the help of the coordinates. In Flickr, people upload over three million geo-tagged images per month [40]. For the organization and search of resources in tagging systems tags are a very important source of information. The quality of the image search is highly dependent on how well the image with keywords, called tags, is annotated [28]. A visual representation of the vocabulary used in these tagging systems is known as tag clouds. To gain a better understanding of the use of tags to obtain, the following will present the so-called tag-space and the associated tagging behaviors of users are examined more closely. The analysis refers mainly to the photo community Flickr and the bookmarking service Del.icio.us. **Figure 2** represents the most popular tags from Flickr in a tag cloud. This type of representation is an alternative to the classical search by text. It allows that users access also information that they have not sought explicitly. They click their way through the tags to images or to others that are similar. The tags in a tag cloud will appear sorted alphabetically. The size of the font depends on the frequency of the tags. Not too surprising is that the terms are chosen very general, since only these are used by most users. Striking here is mainly that some of the keywords differ only in the singular and plural (flower, flowers, or girl, girls) or abbreviations of another (and nyc newyorkcity). To express it only in numbers: there are about 5.5 million Flickr photos tagged with "*nyc*" and about 7.5 million other images are annotated with "*New York*". This means that many images actually reflect the same context, but are not found because they were not indexed consistently. A user writes in a Flickr discussion forum:

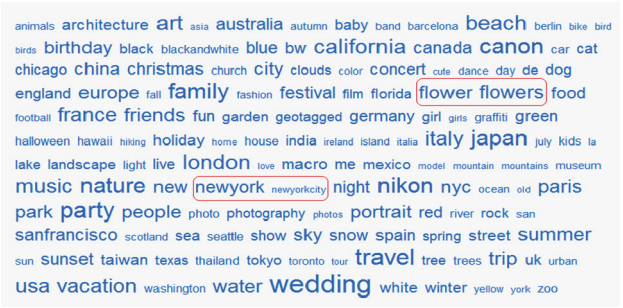



Figure 2. Flickr image tag cloud

"*Is there anything Flickr admins can do about people not tagging their photographs with relevant tags. I'm tired of finding random naked people when searching for baseball shots* " [40]

This raises the question: Are user really tagging in the common interest? The response of another user on it:

"*Tags are for the people applying them, so, although they may have no relevance to you, they may have relevance to the person tagging*" [40]

Users annotate their resources primarily of self-interest. Terms they use may be relevant for them, but in the common interest they are rather irrelevant. An added value to the community arises primarily, if users annotate photos from other users, as they choose in this case rather more objective descriptions.

## VII. ANALYZIS OF TAGGING SYSTEMS

The fact that social tagging is free of ontologies makes it simple for general use but more difficult for machine

evaluation. A good classification (taxonomy) is essential for a large amount of data. The use of a tag of more than one person can provide a common classification scheme. Tags can be recognized as a connection between users and resources. Which users share a tag and what resources were annotated with similar tags is important analytical information in research with folksonomies.

### A. Folksonomies

Users can annotate resources in a tagging system. In the literature this is referred to as collaborative tagging. The collection of tags, created this way is called **folksonomy**. The term "*folksonomy*" consists of the words "*fol*k" and "*taxonomy*" and is attributed to Thomas Vander. Taxonomies are classification systems for data, which are usually hierarchical. Unlike taxonomies, folksonomies have no hierarchical structure and are not developed to purposes of classification, but arise automatically as users tag resources. The advantage of folksonomies is their simplicity, since users have complete freedom in the allocation of tags. There are two types of folksonomies: broad and narrow folksonomies, which is crucial for the analysis of tagging systems.

**Broad folksomony**

In broad folksonomies, many different users (user A to F in **Figure 3**) an index of content creators is made available to any document or similar tags. Thus, the document content from a variety of different or the same subject headings is described [29] . Users that use the same keyword are grouped together in groups and find the document on the basis of this keyword again (arrow in both directions). The most popular Web 2.0 application that uses this kind of multiple annotations is the social bookmarking service Del.icio.us.

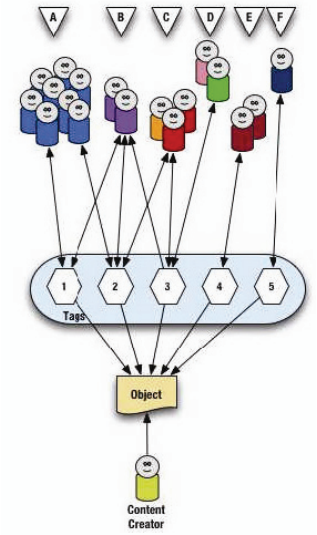



Figure 3. Broad folksonomy [29]

**Narrow folksonomy**

In narrow folksonomies (**Figure 4**) tags are issued only once and recorded only once. Therefore users can only assign new tags. There is no way to count and tag frequencies observed distributions. Mostly the author (or the content creator) creates the first tags. Sometimes it is also allowed to other users to add additional tags. Web 2.0 services that work with narrow folksonomies include Flickr, Technorati, YouTube, a.o..

### B. Weaknesses of Folksonomies

The classification of resources by folksonomy users itself is a problem because the tags are dependable from their own view. This view is understandably subjective, and therefore needs not always to agree with other folksonomy users. This subjectivity limits the retrieval of a resource within the folksonomy. Similarly, ambiguity is problematic for the retrieval of resources because they deteriorate the precision of the keyword search. Here, the precision of the results is enhanced by reducing ambiguous terms and the yield of synonymous words, which were not included in the keyword search. This weakness could be an offset by the use of ontologies. Ontologies enable the creation of semantic relations to represent different levels of abstraction and thus express the relatedness of individual elements. At the same time ontologies allow support for synonyms, homonyms and multilingualism. Ontologies can handle the annotation of resources more efficient, as well as open up extensive search option. Synonyms can be recognized and included in the search: Who is looking for "*Brasil*", is also looking for "*Brazil*". The display of related concepts can guide the search in the right direction: If you are looking for "*mac*", you could also be interested in "*osx*". Upper and sub terms can extend and refine the search: If you are looking for '*newyorkcity*' perhaps in particular for "*central park*" or more generally for "*usa*". Recent research in folksonomies tries to analyze the importance and relationship of keywords. Most of the procedures are based on the co-occurrence of two tags. The calculated co-occurrence value is the number of resources where both tags together occur [30]. We analyzed concepts like the **Actor-Concept-Instance model** and **similarity measures** [25], [49], [50], [51] that derive ontologies out of folksonomies. For detailed information about this topics please see [24]. The insights gained from these concepts will be used in **Section VIII Keyword-oriented group search and ranking in folksonomies** to come up with our own approach for the problem of relevant image search.

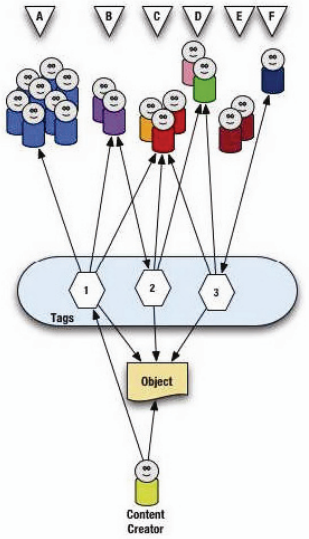



Figure 4. Narrow folksonomy [29]

### C. Quality metrics for folksonomies

The absence of a single controlled vocabulary makes it difficult to assess how the quality of a tag is in relation to the retrieval of the resource. It is believed that the quality of search in tag based systems can be improved if you tag with inter-subjective meaning (a state of affairs for several viewers equally recognizable) or tags that were used by a larger group, determined automatically. This method of analysis, however, is only suitable for systems in which a

term may be given more often. Term frequency within a resource is not allowed in narrow folksonomies. In broad folksonomies they provide important analytical information. The resulting power-law distribution of tags can be used as a basis for the analysis of broad folksonomies. You can concentrate in the search only to the so-called power tags.

"*We hope that offering power tags as a search option improves the precision of search results. We can justify this assumption by the opposing relationship between recall and precision. The one rises, the other falls. In the case of the search only after power tags, the recall - because the entire document-specific "long tail cut off" - is drastically reduced* [29].

In this work term frequency is used for the ranking in folksonomies (see Section C. Flickr groups). In [31] term frequency for the selection of relevant terms are used. Three metrics for the automatic selection of inter-subjective tags are presented for broad folksonomies and tested on a dataset of del.ici.us:

**Metric 1: frequently used tags**
For each tagged resource, the tags are sorted by the number of frequency and the five with the largest occurrence are elected. If a term has been used by several users for a particular resource, this term for an objective description is more relevant.

**Metric 2: tag congruence**
A tag consistency for resource x is defined by the tags that were selected by at least half of the users. This value can be achieved by dividing the number of all different tags for a resource with the number of users. Decisions in various areas of human activities are often made on the basis of the majority. More than half of the people fit in the rule of the majority and often use terms that were already in use. In broad folksonomies tagging may be like a vote for the semantic labeling of a resource [31].

**Metric 3: TF-IRF weighting**
For each tag the Term Frequency Inverse Resource Frequency (TF-weight calculated IRF) is calculated and only the tags by the five highest values are selected. The TF-IRF metric is derived from the term frequency inverse document frequency (TF-IDF). TF-IDF is a standard measure in the field of automatic indexation, to find the best descriptions for documents. When choosing a tag for a particular resource, the TF-IRF formula is taking into account the frequency of keywords for the document. The higher the TF-IDF value, the more valuable is the concept. For the calculation of the TF-IRF value a corpus of similar resources is requires. You get this on by clustering [31] with the Markov Clustering (MCL) algorithm that creates a graph from the co-occurrence of tag pairs. The TF-IRF value can be obtained with the TF-IDF formula conversion [31].

The three metrics were presented to a record of del.ici.us tested with 3.4 million users from different bookmarks from 30,000 in 2007. Then the test with an online survey was bound to find the most appropriate metric. **Metric 1 (frequently used tags)** provided the best results [31].

**Evaluation of the approaches**
Most of the described approaches and ideas mainly work with co-occurrence, simple clustering algorithms or the vector space models. The resulting similarity values can serve as a basis for a similarity graph. In the Actor-Concept-Instance model resources, users and tags are represented as nodes. For the relationships of the tags are only the connection graph "*user tag*" and "*tag resource*" decisive. The former provides an ontology based on users with similar tagging behavior and the latter an ontology annotated on similar objects. This type of graphical modeling of tagging systems is an important basis for ranking systems, such as the **FolkRank** [32]. The algorithm is based on the idea of the PageRank algorithm and is used for the ranking in folksonomies. The PageRank algorithm computes rankings on node with the idea that a node is important if many other important nodes point to this node. Based on the FolkRank algorithm, this means that a resource is then important if it is connected with important users or tags. The FolkRank is a modification of the PageRank algorithm, as this cannot be applied directly on folksonomies. The FolkRank algorithm determines a lot of relevant resources and users for a tag. This information can be used to assist the user in the annotation and in the search.

The view of each user on a resource is subjective. Many resources (images) are ambiguous and are therefore interpreted differently by different users. The degree of content development is crucial. Some users use more general terms such as "*animal*", while others are more specific such as "*dog*" or "*puppy*" that complicate the search of resources. Users can also describe the same or very similar pictures with different keywords. While a user an image with "*lake*" annotated, this may be another tag with "*sea*". This problem is to use the surrounding to identify tags that are based on co-occurrence relationships. The **co-occurrence relationship** is highly dependent on the amount of data. For a very large amount of data (like Flickr), it is relative, since one in very many different resources for two very similar tags like "*animals*" and "*animal*" can have a low similarity value of 0.06. The main reason is that usually the tags are assigned mainly to Flickr only by the creator and he did not worry about the plural, singular or synonyms. An image that was tagged as "*car*" will probably not be additionally annotated with "*automobile*" by the same user. There are 10 times more images in Flickr tagged with "*car*" rather than "*automobile*". Procedures that try to build a threshold value from the tags top and narrower relations have the problem that many special tags are collected as generalized tags. Therefore, these are suitable only for supporting the user in his choice of auto tags and less for an annotation. Another important factor is the multilingualism. Members use many resources for different languages. Usually the mother tongue is combined with English. In addition, users can annotate a resource from different backgrounds together what is an advantage for the general search, but it brings considerable

problems for machine evaluation. Many tags mean the same thing but because of different languages they have a small co-occurrence and reduce the effect of similarity calculation further. It is difficult to reach a clear classification of the tags solely on the information of the co-occurrence frequency and the frequency of tags in a library. The co-occurrence frequency allows that the less descriptive tags (which are rarely used) are eliminated.

The approach to combine folksonomies with existing ontologies provides lightweight ontologies. It filters the irrelevant tags and finds relationships between relevant concepts. The problem of ambiguity can be minimized over the Semantic Web ontologies. Considering the enormous amount of data which e.g., Flickr provides (over 2 million images per day), this is too complicated, but for limited amounts of data very demanding. The approach adopted here identifies the relationship between tag pairs on the semantic search engine Swoogle that has only the English language. Because tags are often multilingual, this approach is suitable for images that are tagged in English only, and is less effective for multi-language terms. This problem could be limited, if we automatically translate any foreign tag into English. Such an application is presented in [33]. It translates the search terms automatically in up to six different languages. In combination we can get multilingual image retrieval from Flickr.

Quality metrics for folksonomies are suitable for the selection of relevant tags very well. Unfortunately, these mainly take into account the term frequency applicable only in broad folksonomies. Narrow folksonomies cannot show certain frequency distributions of tags since all tags are equal (all tags come only once). Therefore, the presented metrics work only for broad folksonomies. A direct application to narrow folksonomies does not provide the desired effect.

The indirect concept is the **gradation of the tags** within a tag list – so we can develop other methods to determine the **relevance of resources' tags**. One solution for this is presented in Section VIII D.

The relevance of the assigned tags is critical for the retrieval of the images. We presented some approaches that examine the relevance of keywords. Since the quality of tags is dependent on the co-occurrence relationship and therefore on the tagging people the similarity graph is an efficient modeling method for folksonomies. This helps to consider the tagging behavior of users, the co-occurrence and term frequency simultaneously. Unfortunately, this information alone is not enough to improve the quality (relevance) of the tags automatically. But the information is well suited to support systems in proposing tags to the user. Some approaches attempt to get additional help by external sources such as Wordnet , Wikipedia, Google or the Semantic Web search engine Swoogle. These make it possible to find a genuine search for synonyms or discovered ontologies. Synonyms help eliminate the significance of ambiguous tags.

In the next Section we introduce our own derived ideas and approaches to allow **image search optimization in folksonomies**. For experimental purposes only we use the Flickr online photo community. Flickr provides next to the API and the tags other metadata such as description of images, comments and number of clicks (views). This information can be used to make a statement about the quality of the found images.

## VIII. KEYWORD-ORIENTED GROUP SEARCH AND RANKING IN FOLKSONOMIES

Groups allow pre-selected content and increase the precision and relevance of the recall. Our idea to improve search results is a keyword-oriented group search and ranking. We developed a tag ranking game called qRANK to rate and rank Web resources. Flickr allows its users to organize pictures in groups and related groups in collections. Groups, tags, views and comments are important information to learn from folksonomies. The aim of this work is not to develop a global algorithm for the complex search problem in folksonomies. Rather, we implemented and evaluated ideas and methods to optimize photo relevance and quality for Web photo searches. A methodology which allows an automatic classification and ranking of photos of their attractiveness was developed in [35]. Photo attractiveness is a very subjective term that depends on many factors. The feedback from the user will supply important information for classification and regression models to create, based on visual characteristics of images and the metadata

„*In a wider system context, such techniques can be useful to enhance ranking functions for photo search, and, more generally, to complement mining and retrieval methods based on text, other meta data and social dimensions.*“ [35]

Visual features such as "*color*", "*contrast*" and "*rudeness*" of images and other metadata such as tags and favorites lists are examined. The combination of visual and textual features yielded the best results for the ranking according to a photo's attractiveness.

Here the main issue is the quality of the image search. The quality of a search result is determined by the intention of the searcher. Therefore, it is an advantage to consider the search behavior and motivation of the user precisely. In general, a user has the following interests:

1. **Precise search**: the user is looking for a specific image or images for example of the Eiffel Tower.
2. **Search topics**: he is looking for a picture or pictures on a specific topic such as only black cats or dogs of a particular race.
3. He has **no particular intention** of looking rather out of curiosity and wants a closer look at village (vicinity search).

### A. Attractiveness of pictures

This approach should help to determine the precision of the images by the attractiveness and popularity of the photos. A scenario for an exact search might look like this: A user searches for a picture of the new city hall in Hanover to use

Figure 5. Flickr standard search for the terms „Rathaus“ and „Hannover“

in his school lecture. He used the two keywords "*Rathaus*", and "*Hannover*". Therefore the **standard keyword-search in Flickr** provides 175 results. We can display the first ten images at random and get the following pictures as seen in **Figure 5**. Also there are some images on the town hall, none of this is what he really wants to use for his work. Of course, among the 175 photos found there are some that correspond to his ideas and with a little patience he would find the right image. However, the user wants to find the photo that is relevant to his search as soon as possible. The relevance of the image here refers to the given information content for the user, as generally all images may be relevant. The intent of the user (use: seminar work) implies that the content of the image must satisfy the search term clearly. Relevance is indeed a relationship between an image and a user. A tag and a picture are defined as relevant, if the tag only describes aspects of the visual content of an image [36]. In the course of this work we call **relevance** (also used in precision) the degree to which the content of an image corresponds to the entered search criteria. This degree of precision can be used to classify images. Besides the problem that many images cannot be found because they were annotated with little or inaccurate tags, there is a further problem, to assess the degree of relevance. For some queries you get a very large selection of Flickr images that are different relevant. Since one is usually interested only up to a fraction of these images, a ranking of the found images is required. There is a patent publication of Yahoo! for Flickr which deals with this problem [29]. There are set five criteria for a ranking by interestingness in narrow folksonomies:

1. The number of tags to a document
2. The number of people tagging a document
3. The number of users that get the document after search
4. The relevance of the tags
5. The time (the older the document, the less relevant)

Most of these criteria are closely related. The first two criteria are important for the relevance of the tags. If multiple users annotate an image with different terms, they create a multidimensional view upon the resource. Suitably chosen tags facilitate the search. If the terms are very different, the search is inaccurate. An image that was tagged by different users reflects also the popularity of this picture again. Photos which are described with many tags are found more often. The criterion of time is not applicable, because a picture does not lose its relevance over time. The feature "*Interestingness*" is described in Flickr [40] as follows:

"*Many factors affect whether something is on Flickr interesting (or not). It depends on the origin of the clicks, who commented when the image of who identifies it as a favorite, which tags are used, and many more factors that change constantly.*"

As the components are related is deliberately not discussed deeply. Derived from [29] we define three different sets of criteria for the ranking in tagged documents (see **Figure 6**) which are of importance for our work.

The first volume contains procedures that relate to the semantics of the tags. The relevance of the tags can be determined using the method presented in the previous section as the TF-IDF weighting, the cosine similarity or the FolkRank algorithm. In addition to these criteria, there are other factors, such as click-through rates, the number of comments and favorites list, which can be crucial to a relevant search (collaboration). In addition, you can include the relevance of terms, the feedback of the users with (prosumer). This can be done in a question-answer game where users assess metadata of resources playfully.

For a relevant search, some of the investigated options shown in **Figure 6** are examined. In the next approach, we use the **click-through rates** and the **upload date** of the pictures and would like to examine whether images, which are often viewed at the same time have a higher relevance. About the interface of the Flickr API can about each picture about click rates (views) and the upload date to be fetched. The number of clicks is an implicit relevance feedback, "*they are in a high degree collaboration-oriented ranking criterion in the sense of Web 2.0*" [29]. The mark as a favorite reflects the attraction and popularity of the image. In general, one can assume that with increasing click rate, the favorite rate rises. Thus, we extended our search with an additional function that sorts the pictures by clicking the spending rate. The click rate is a picture of the dependent "*upload date*" dependent. Photos that are longer online have generally a

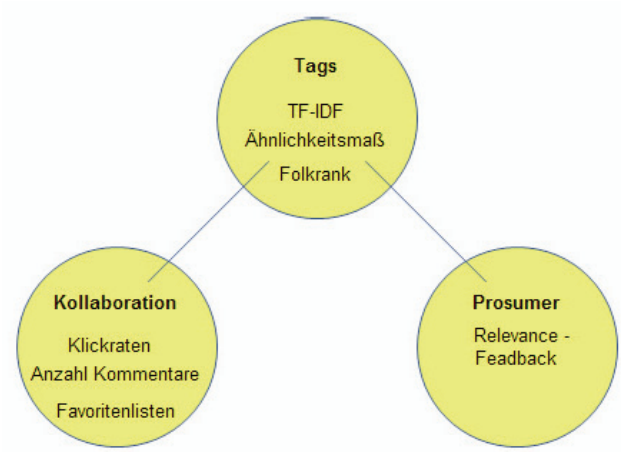



Figure 6. Ranking criteria in folksonomies

Figure 7. Extended Flickr search for the terms „*Rathaus Hannover*“ with precision formula

higher click rate than actual pictures. To counteract this, the upload time in the calculation considered. This function is called the **precision formula**, resulting from the division of the click rate and the time (in seconds) that a picture is already online sets together. About combines the precision of the ranking formula for "interestingness", the relevance of the retrieval set is clearly improved. The same search from the previous example, sorted according to the precision value, returns the data shown in **Figure 7** with the first ten images that the user receives after a search for "*Rathaus Hannover*". The weakness of this method is that the images are very new, get assigned a higher weight than older ones. An image that has ten clicks on the first day would have a very high precision value without being necessarily relevant for our search. The click through rate alone is not an absolute indicator of the relevance of a search. The click-through rate of an image rather reflects their popularity again. This in turn depends on several factors. As a rule, to Flickr photos that belong to a broad community often looked at. Images that contain many groups, and its creator are linked with many other users have generally higher click rates. This means that the pictures were annotated and rather inappropriate for a subject search are not relevant, but can have a very high popularity. In Section VIII C., an approach is presented, how images have grown to their relevance.

A major problem in the search for relevant images is the ambiguity of the tags. The tag "*Paris*" can mean a city in France or a city in the U.S. or even refer to a name. When the user searches the tag "*Paris*" for pictures of the French capital, he will receive, among other things pictures from America or from people who are called Paris. This can reduce it but if we expanded the query with related terms. In the research of folksonomies this approach is the "*tag of suggestion*" [37] or called tag recommendation [28] [30] and can be used for two things. First, you can use it to help the users to support the annotation. Recommendations will help users to clarify the image content as well as reminding them of related semantics which may otherwise be ignored [28]. On the other hand we can extend the inquiry with other tags in order to achieve a more relevant search. We concentrate here on the second.

### *B. Tag suggestion*

The idea of tag suggestion is used in this section to specify the search for images. From previous considerations we know that tags are ambiguous, imprecise and often irrelevant. Linguistic differences and the fact that users are not professional tagger make it difficult to find the pictures in Flickr. If a user has annotated a picture with the words "*cat*", "*white*" and "*charly*" we will not find this picture, if we search for the keyword "*Katze*" (German translation). In Flickr, there are twice as many images that are tagged as "*cat*" than with "*Katze*" and also about the same as many pictures that are tagged with "*cats*" instead of "*cat*". Even if these images actually reflect the same content, they form different result sets in Flickr. Some works in the tag list folksonomies combine an image with relevant concepts from other sources such as WordNet [36]. In this paper, we focus primarily on the query and try to isolate the problem of imprecise tagging, as we show related tags to the user automatically. Here the question is expanded by the user with the selected terms. Based on the above example, the user gets a list of related tags containing terms like “*cat*” and “*cats*” while searching for "*Katze*". These are terms that often occur together with the search word (co-occurrence relationships). On extending the search to several terms, also increases the amount of results.

The query extension can be used to further narrow down the search space. This is e.g., in qMAP used to reduce the problem of synonyms. If a user searches for the word "*apple*" searches, it is not clear whether this term refers to the fruit "*apple*" or to the company "*Apple*". Such an inquiry would yield many irrelevant images. However, if the request is extended with an additional term such as "*fruit*" or "*Mac*", then its ambiguity is eliminated. In this simple case, the searcher possibly finds out on his own that his request is not clear and would change or expand his search with a further term. In most cases, however, a user does not worry about whether his chosen search term is ambiguous and much less he finds an appropriate term with which he can formulate his question precisely. An improperly selected tag means that the results are again irrelevant or relevant images are not found. A selection of tags that are related to the term used by the user in a strong correlation facilitates the search. In qMAP, the user gets a list of related tags available for selection like in the query extension. The terms selected by the user are involved in the request and only images are displayed that contain the tag list and all of the keywords. A multi-query search is also suitable for general subject searches: A user searches for a specific topic such as black cats. This is the request for "*cat*" extended with the term "*black*" and searched for images that contain both words. In response, the user gets only pictures that at least contain the two concepts “*cat*” and “*black*”. For a more precise topic search, this version is less suitable. From the knowledge that many images are annotated inaccurate, it can be assumed that the method of query expansion also provides images that do not contain any black cats. On the other hand, there are also pictures that would have been useful to the user on the context, but are not found due to the lack of tags. The

number of tags per image is very limited in Flickr [40]. This is because most of the pictures are annotated only by the creator and are not tagged with many words. In addition, a user does not take the time to worry about and to discuss alternative and more detailed tags. In contrast, the **groups at Flickr** are used more often. A study in [38] has found out that over the half of the users (about 8 million) share at least one Flickr photo with a group. Flickr groups are self-organized communities with common interests [38]. The analysis of Flickr groups is an important step to find relevant images that were inaccurate or not tagged. In this study, the groups are used primarily for the **subject search**.

*C. Flickr groups*

A group is a collection of people and objects that are either in physical proximity or share certain abstract properties. The main goal of a group is to facilitate the exchange of resources in a community. In contrast to the similarity graph in previous sections, groups are not generated algorithmically. They arise spontaneously, not by chance:

"*Users participate in groups by sharing and commenting on photos, most often on specific topics or themes, like a popular event, location, or photographic style.*" [38]

Such collective behavior modes offer alternative ways to understand and analyze visual content. Grouping is a simple and well-received folksonomy function, which provides valuable information to detect relevant resources and improves the quality of the search [41]. Most groups had a clear theme, and are sorted in this context issues.

"*Two images are similar if they belong to the same Flickr group*" [47].

Users who are involved usually have the same interests. They exchange information and knowledge by group discussions and comments about the pictures. The resulting **collective intelligence** enables that the images are better annotated in well moderated groups. Members, who are friends with each other, develop similar approaches to an image. In [42] the effect of the grouping in a tagging system is presented with **Group Me!,** in which the user can organize any resources from other tagging systems in groups via drag-and-drop. Group Me! allows not only tagging of resources but also tagging of the groups themselves. The annotation of resources can always be considered in the context of a particular group. This provides additional relationships that can be used for the quality of the resource ranking:

"*Tagging resources is always done in context of a certain group. This group context gains new relations between entities of the GroupMe! folksonomy, which consists of user-tag-resource-group bindings, e.g., the group's tags are likely to be relevant for the members of the group, and vice versa. Such new relations enable advanced folksonomy-based ranking strategy.*" [43]

A ranking algorithm is in Group Me! presented that uses the effect of the grouping for the ranking in folksonomies. The "*Grank*" algorithm based on FolkRank returns through the use of the group structure better results than the general FolkRank algorithm [43].

In Flickr, groups are collections of people who join voluntarily in a community. The collections of resources that are collected by the group members are called "*group pool*". Each user can create any number of groups. There are three different types of groups that are crucial to the search for these:

(1) public, everyone can see the group photos and join the group.
(2) public, everyone can see the pictures, membership by invitation only.
(3) private, no one can find the group, membership by invitation only. here consider only publicly accessible groups.

Here, we concentrate on public groups only. In [38], the group structure of Flickr is analyzed. The average number of members per group is approximately 317 (**Figure 8**).

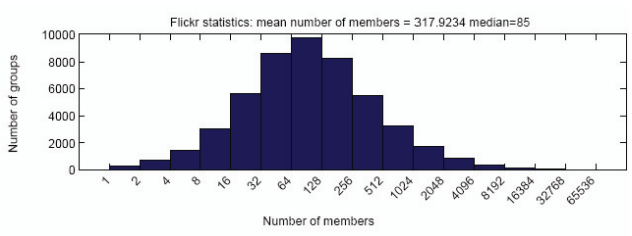



Figure 8. Analysis Flickr groups "number of members" [38]

Unfortunately, there are also many groups in Flickr with very few members and even groups without images. These provide no information in this work and are known as "*spam groups*". The average number of photos in a group is

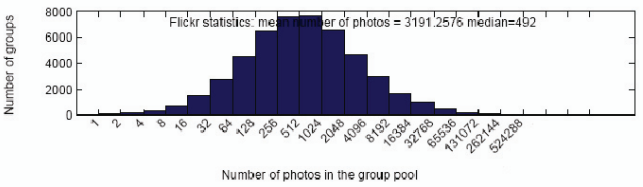



Figure 9. Analysis Flickr groups "total images" [38]

approximately 3191 photos (**Figure 9**). Both images are a proof that the **exchange of photos in groups** is an **important activity among Flickr users**. More than 50% of the users share at least one picture with a group. Over 25% of the members share at least 50 images [38]. A photo can also be included in several groups. Groups ensure a higher exposure of the photos. They offer the user a wide selection of relevant images for a specific topic and make the photos easier to find. Similar difficult to the search for images is the search for relevant groups:

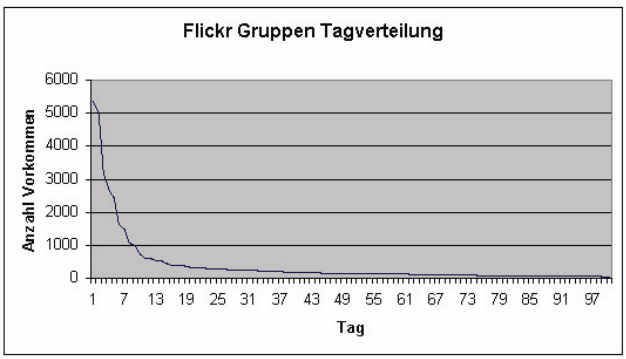



Figure 10. Example tag distribution in a Flickr group

"*In practice, finding groups on Flickr is relatively cumbersome and does not make use of the plethora of meta-data available in the user groups and photo collections*" [38].

Groups are found in Flickr in the first place through their group name or description. The title of a group is not always perfectly. The description is often too broad and not specific enough and we find irrelevant and too many groups for a specific topic. According to [38] 60% of the groups consist out of one to five relevant subjects and only in 10% of the groups we find more than ten subjects. Unlike in Group Me!, users can annotate only the pictures in Flickr. The number of tags in a group is therefore limited by the maximum of 75 words the images can be described with. **Figure 10** shows the 100 most used tags in a group with a total of 15.222 elements. At the beginning of the curve a few tags are placed with high values, the right end is composed of many nearly equivalent tags. This type of distribution that is similar to a **power law curve**, was discovered in broad folksonomies by Thomas Vander Wal [29].

The tag distribution of the Flickr groups is almost identical with the ideal power law function. The green area in **Figure 11** contains tags that are found in most resources. These reflect the collective opinion of the group members and are more relevant for the groups' subject. In the yellow area, includes the so-called "*Long Tail*" as the special tags. These are rather subjective tags that are related less to the subject in the group. There are no annotated Flickr groups, so one can derive the tags of the images to the groups when considering the groups as one resource. The tags, which occur frequently, are more relevant to the topic in the group. For further and detailed information about our group mechanics please see [24].

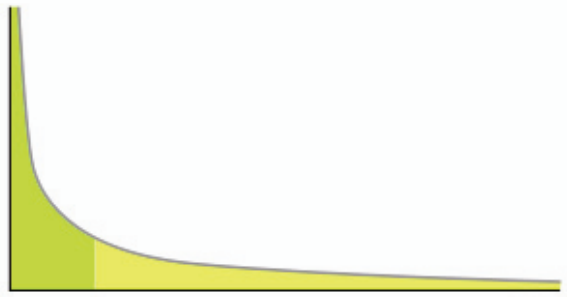


Figure 11. Power-Law curve [29]

From the previous considerations we now deduce our **tag-based search and ranking procedure for Flickr groups**. The approach builds on the search methods used in Flickr, but then considers ranking of the search results by the most used tags in each group. In addition, this method eliminates groups that have little or no elements. For the ranking of the groups following information should be considered:

1.) The members and the number of elements.
2.) The most used tags with a weighting factor.
3.) The titles and the descriptions of the groups.

The idea is that groups that contain most of the pictures in the ratio for the given tag are most relevant for a subject search. Since the groups are primarily used to get the most images on a specific subject, only groups are interesting, that provide a certain amount of images. Therefore, the group ranking process ranks the groups according to the quantity of images that are annotated with the wanted keyword. The most commonly used tags are elected as representatives of the groups.

**Application flow**

First, the search term is compared with the most popular tags in a group. All groups that contain the search term as tag are weighted on the frequency of their tags. If we look for groups that follow a clear theme, then the weighting is based on the number of elements with this tagged term divided by all the elements. If one is interested in the most pictures to a search term, then the occurrence of this term is used as a weighting factor. If we got the group with the most appropriate images, we can do a keyword search within this group and for example sort the images according to their relevance with qRANK (see Section IX).

Then we successively take into account the following criteria:

1. The compliance of the users' search term with the groups' tags is examined.

Since users are using a known way in their annotation usually and not all forms of a term together, the above condition is extended. A user who searches for "*church*" is also interested in pictures annotated with "*churches"* and "*Kirchen*".

1.1. An English translation of the search term is taken into account in the search
1.2. To see the similarity between the plural and singular, the Levenshtein metric is applied with a distance of two.

The Levenshtein metric can be applied easily, because we can usually consider, that terms like "*Church*" and "*cherry*"

| Rang | Rang des gesuchten Begriffs innerhalb der Gruppe | Mitgl. | Bilder | Top-Fünf-Tags | Bilder mit Kirche/ church | |
|---|---|---|---|---|---|---|
| 1 | 4 | 24 | 123 | münchen munich architecture kirche church | 77 | 1 |
| 2 | 0 | 1 | 0 | keine Tags | 0 | 11 |
| 3 | 1 | 2 | 1 | kirche parchim | 1 | 10 |
| 4 | 0 | 1 | 51 | judith thomas moe hochzeit | 0 | 9 |
| 8 | 49 | 420 | 2358 | jesus christianity hymn chant christ | 143 | 8 |
| 9 | 135 | 6671 | 79863 | italia italy anticando church roma | 14598 | 7 |
| 47 | 5 | 1643 | 23796 | church europe cathedral architecture kirche | 11447 | 4 |
| 138 | 2 | 560 | 30354 | church kirche carving austria österreich | 8433 | 6 |
| 245 | 4 | 91 | 2173 | church europe cathedral kirche architecture | 1280 | 2 |
| 367 | 3 | 180 | 1441 | gothic architecture church cathedral england | 419 | 5 |
| 448 | 4 | 1235 | 10809 | church architecture europe kirche cathedral | 5217 | 3 |

Figure 12. Flickr group analyzes for the term "*Kirche*"

are different in more than two places. They are not in a singular-plural relationship and are not together amongst the most used tags found in a group because they represent two very different topics.

If a query matches with one of the top five tags, the affected groups are ranked according to the weighting factor. If several terms match the sum of all weights is formed. If the tag list of a group does not contain the search term or empty groups are weighted with Zero. All groups that are equally weighted are ranked according to a second criterion: the number of images. If the number of images is also equal, as third the number of members is taken into account. As a result of the procedure we get a ranked list of the groups. This method is especially effective if we seek for general subjects that provide a wide range of groups. **Figure 12** contains an example part of the list of results for the term "*church*", which provides a total of 1551 groups. Since the list in fact, very long, here is shown just a snippet. The column "*Rank*" in the table gives the position in the list that Flickr (sorted by the relevance) returns. The idea of this group ranking procedure is to **find the group with the most relevant images**. The red numbers in the table represent the rank that our presented method derived. At the first rank position, both lists are still identical, but the remaining positions differ massively. Many groups which "*Kirche*" in their top five tags are weighted stronger by Flickr than groups that use the tag "*Kirche*" not at all or very rare. The explicit consideration of the tags' plural/singular and the inclusion of the terms' English words come to significantly better results than the standard Flickr search. Since Flickr does not provide intentionally the needed data for the approach, they must first be created. At once, Flickr allows only a maximum of 500 pictures or information per request to download. In order to realize a dynamic and non-redundant storage concept, the idea of the Actor-Concept-Instance model has been implemented. For further detailed implementation details please see [24].

### *D. Tag ranking*

The approach discussed in the previous section allows ranking the groups according to their relevance. Only term frequencies will be considered, which are calculated from the tags of the images. The images in the groups are not ranked yet. In the following, the idea for an image ranking game called **qRANK** is presented. It provides important information to **rank the tag list of an image automatically**. This information is then used to sort images according to their relevance.

Narrow folksonomies like Flickr, have a major disadvantage that they do not allow the frequency distribution of the indexed terms. Therefore, it is not possible to observe tags abundances and distributions within a resource. All tags come only once, so that we do not have simple methods to distinguish between relevant and irrelevant tags. A user can tag his pictures in Flickr with up to 75 keywords. In general, the tags are chosen arbitrarily.

With known methods we mentioned in **Section VIII A.** like TF-IDF weighting we could determine the relevance more precisely automatically. Here we like to introduce the **different approach** qRANK, which allows us to classify the tag list of an image in a **game-based** way. This game should investigate in how far the process of the players acquired knowledge in a dynamic ranking may change the tag lists quality. With each pass of the game improved the tag of an image that can be used for further analysis, particularly for the improvement of the search list.

## IX. *qRANK*: A TAG RANKING GAME

Most of the analysis so far considered folksonomies that deal mainly with broad folksonomies. The resulting frequency distribution of tags examined is an important indicator to determine the relevance of one tag in reference to the description ability for a resource. This collective knowledge can provide a statement about the relevance of a tag. The implementation of the tag rankings (previous section) by **a game that implements the idea of the power-**

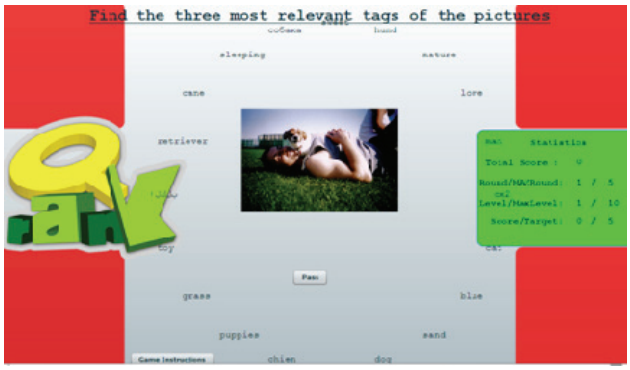



Figure 13: qRANK interface

**law curve** would provide **additional information** for the ranking of images. Most approaches to rank folksonomies are based much more on the FolkRank algorithm [32] or ranking techniques based on particularly elaborate calculations [33]. In this work the pictures' tag list is sorted according to the relevance of their tags. At the same time the tag list is extended and annotated with **new valuable terms**. qRANK (see screenshot **Figure 13**) queries available Web services (almost RESTful) and embeds returned content in a predefined gaming setting. Here we added some algorithms to enhance the precision (relevance) of the search results like e.g., interestingness rating or precision formulas for folksonomies. Additionally, every gaming interaction is logged and ranks played content enabling the users' collective intelligence by and by. Results are stored in qKAI but are still semantically interlinked with the provenance source not to lose the resources' context and for updating. Techniques used are semantically Linked Data (annotation, interlinking), server-side Java, Adobe Flex/Flash and a MySql database – to be flexible in representation. For further implementation details please see [24] and [48].

### *A. qRANK: game description*

The user gets presented a picture and a list of twenty tags. His task is to choose the three most relevant tags that reflect the subject of the picture best in his opinion. Subsequently the chosen terms are reviewed by the rank in another list, and rewarded with points depending on the tags' rank position. For each term that is included among the top five tags, the player gets three points. In positions six to ten the user gets two points and for the positions 11-20 he receives one point. If the term is not included in the list or the rank is below 20, the user gets no points. The motivation of the player is to achieve the maximum number of points per round to get to the next level. The game consists of ten levels. In each level the player gets five consecutive images displayed and can reach a maximum of 45 points. The barrier from level one to two is at 20 points, and increases for each level by 5 points. So from level 6 you only come further to the next level having full points.

### *B. qRANK: architecture and backend*

**Figure 14** describes the components and the approximate sequence of qRANK. We downloaded a data set of relevant images to a certain topic from the Flickr web service and stored it in a MySQL database. The information for all the images are recorded in one table. In addition, the related tags that fit best on this subject are saved in another table. In the third table (image tag list) all tag lists of the images are managed. The image tag list consists of the terms that users have used to describe this picture in Flickr. A fourth table (ranked tag list) is filled dynamically. This is filled at the creation of the game with ten terms of the actual image and a related tag list tag. The ranked tag list contains for each term a counter, which is used to count the frequency of the term.

By chance, the player gets presented a photo and 20 matching tags. The tags will be selected for a specific principle from the tables "*related tag list*", "*image tag list*" and "*ranked tag list"*. This achieves a useful combination of tags. In the very first run of a picture the length of the "*ranked tag list*" is set to twenty. While producing the amount of data every tag list will be employed with ten randomly selected tags out of the "*related tag list*" and "*image tag list*". The number of tags in Flickr images is

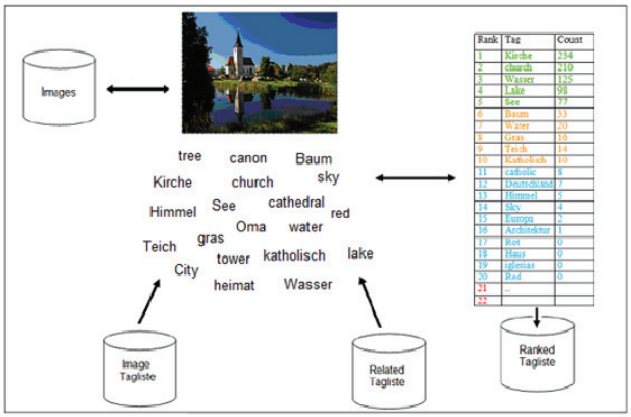



Figure 14. qRANK concept with tag lists

different; many images have less than three tags [26]. If a picture does not have ten tags, so in this case, the missing tags are added from the related tag list. These twenty tags are then stored in table "*ranked tag list*" and build the new tag list of images that is sorted dynamically through the game.

### *C. qRANK: gameplay*

After a player has selected three terms, they are compared with the tag list and awarded with points. Since the first run of the counter of tags is to zero, an additional condition is defined: If the counter of all terms is the same, the player gets his choice irrespective of the maximum score for that round. From the second pass (for each image) is the selection tag list combined out of the first 10 tags of the ranked tag list with 5 randomly selected tags from the "*related tag list*" and the actual "*image tag list*". With the selection of the top ten ranked tags from the tag list, we ensure that terms that are more relevant are selected with a higher probability. Even here, it may happen that the actual tag list (image tag list) contains less than five tags. In this case, the remaining tags from the ranked tag list are added. To prevent duplicated tags, the randomly selected tags are compared with the related tag list and the actual tag list of the images with the first ten terms from the ranked tag list. The logic of the game is developed as web services. To get a better overview of the game's flow from the perspective of the player, it is described as follows:

1. The player gets a random image and a collection of unsorted tags. He has to choose the most relevant three terms.
2. The chosen three terms will be compared with the ranked tag list.

2.1. If they match, he gets (depending on rank of the term) points and the counter of the tag are

incremented.

2.2. If the selected tag is not included in the ranked tag list, this is added thereto and the counter is set to "1". The player gets no points. This ensures that the tag list is ranked and expanded with additional terms. A limit on the maximum number of tags is not set in the game.

However, the maximum number of tags is fixed by the quantity of the tag list and the list of related terms. The primary objective of this game is to **evaluate the information gained from the existing tags to an image**. Through an extra box users can also add optional new tags.

### *D. qRANK: ranking of the images*

The information, which is calculated from qRANK can easily be converted into a ranking of the images. Therefore, qRANK itself is already a precursor of the ranking. The more an image is played, the more meaningful is the tag list. The idea behind this ranking is similar to the group ranking. A picture is evaluated collaboratively and as a result we gain a weighted list of objective tags. The subjective tags that insignificant for information retrieval fall out automatically. Tags that do not explicitly describe the content of an image and only have a meaning for the person, who assigned them, are not included by the public (the players). The result is a tag list for each image sorted by relevance. The degree of relevance of a term for an image depends on the objective consideration of all persons who have played this picture.

The result is the basic principle of this tag ranking process. In this procedure, any tag from the ranked tag list, which belongs to the image, is weighted. The weighting consists of the simple calculation of the number of times this tag was chosen, divided by the sum of the possibilities that he stood for selection. The relevance results here out of the tag's selection counter in relation to all other tags' selection counters. A valuable statement is possible if an image is played with certain frequency.

## X. EVALUATION

We have seen that the search for relevant groups and image in folksonomies represents a fundamental problem. Some related approaches have been described in this paper trying to use the resources metadata (tags to classify). From the analysis of these approaches in this work, new ideas have emerged, which were implemented as a prototype. In this section the effect of the implemented approaches in this work to search for relevant groups and pictures are shown. The experiments described below compare the **standard keyword search in Flickr** with **our group ranking method** and our **game-based approach** (qRANK).

### *A.* Experiment 1: group ranking

The aim of the group ranking procedure is to find the group with the most relevant photos according to a topic or term. These are the groups sorted by relevance to the topic. To compare the method with the search for relevant groups in Flickr, we stored the term "*Kirche*" of 100 groups with information about the images, tags and users in a MySql database. The groups search on this term has found 1640 groups at the time of the experiment. To download all the required information over the Flickr API, we have to provide several queries for one group. Unfortunately, the Flickr API does not offer the function to determine the occurrence of a specific identifying tag at the time of this work. Therefore, an additional methodology was created to determine the frequency distribution of tags within a group. 100 groups have been considered demonstratively here, with their 100 most used tags. A data set of a million images and over 100 thousand emerged out of this. To optimize the performance of the database query the set of tags was reduced to 100 most used tags per group. The groups are selected as follows: Fifty of the groups are also the first 50, as they are returned by Flickr and the other half, randomly selected groups from the rest of the crowd.

### *B. Result experiment 1*

**Figure 15** represents the number of relevant images of the first 20 groups that Flickr [40] provides on the query "*Kirche*", compared with the process of this work. The red bars describe the results from Flickr and the green bars, the results with the group rankings from this work. During the first eight groups in Flickr together provide a total **of 100 images** to the search term, with the groups ranking procedure we get in the first position a group with **5262 image**s. Considering that Flickr has all of its data available and here we included only 100 groups, the procedure becomes even more important. As we know Flickr does not explicitly take into account the tags and still less the number of images as a relevance criterion. Therefore, seven of the first eight groups in Figure 15 are empty, while the groups ranking procedure sorts the results by the number of relevant images. The relevance of the images is judged here by the strong commitment of the Flickr groups. The relevance of a group is not necessarily dependent on the number of matching images in a group. A group with fewer elements could well have more relevant images as one with more pictures. In this case, we can optimize the groups ranking method by combining it with qRANK. Thus, the tags of the images are evaluated within the groups by qRANK and the weight is derived based on the evaluation of the tags for the

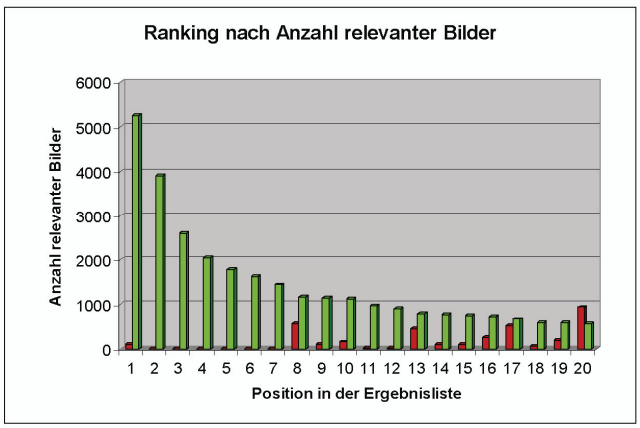



Figure 15. Results of the Flickr group rating approach

image and the group.

### C. Experiment 2: game-based picture ranking with qRANK

For this experiment, we put two different versions of qRANK online for one week. The first version consisted of 250 randomly selected images to the topic "*Kirche*" and the second version of 100 images selected specifically on the topic of "*Hund*". The users should select the most relevant three terms for the image. In the first scenario a user always had to choose one of the words even if he is not sure in his choice. In the second game, the user could press a pass button to get the next picture if he found no suitable definition.

### D. Result experiment 2

The first variant of the qRANK was at this time not played as often as originally expected, so that no term was selected more often than twice. This value was too small to be a statement about the relevance of a tag. The second variant of qRANK was played more often and provided due to the small amount of data desirable results. An evaluation of the ranked tag list of the one hundred pictures provided, showed that 56 of the pictures had their most relevant tags in the first place. Only nine pictures did not have their most used tags in the first four positions (see **Figure 16**):

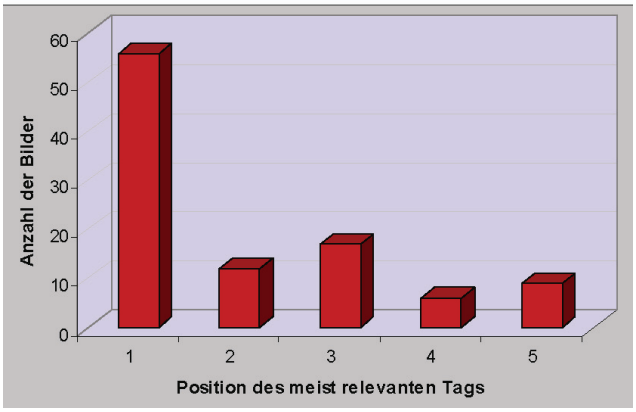



Figure 16. Positions of the most relevant tags

A one-week game period brought the result that **91% of the images that were played** during this time, had their most relevant tags **to the first four positions**. These results illustrate the effect of the approach. The aim of this experiment is not necessarily to find new terms for an image, but to assess the relevance of the existing tags depending on the content of the image. To make a useful statement only images were considered, that were selected at least four times. Striking here was that users often choose terms in different languages or plural/singular relation. So many images in the ranked tag list appeared often in different languages. Regarding the search process, this is not necessarily a disadvantage, since users search more multilingual. For the evaluation of the concepts in qRANK it is disadvantageous in the long run, as these terms are more preferred, and thus reduce the probability that other terms are selected. This problem can be limited, if we determine these relationships before automatically.

### E. Resume

All over, information quality enhancement is getting more and more important – especially regarding the flood of autonomous Web resources without responding authorship. We presented exemplary the role of information quality in web-based information and knowledge transfer with smart interaction.

We adapted an existing assessment model to our purpose in qKAI and showed some examples for enhanced, rating-based interaction that is suitable to qualify Open Content stepwise in an incentive way. Incentive for user participation and interaction is implemented in qKAI as game-oriented, ontology-based and global rewarding model for any kind of interaction. Information quality can be utilized as a tool to derive personalization and user preferences in web-based information and knowledge systems, because it offers a.o. metrics to determine the fitness for use of autonomous, distributed resources.

The evaluation of our group-ranking and the game-based assessing approach for Flickr images showed promising results and the contents' quality increased obviously. Single tasks are reusable and combinable in different scenarios (implemented as atomic Web services).

## XI. FURTHER USE CASES AND EXAMPLES

### A. qMAP: A geo-coded visualization of Open Content

With qMAP [24] we implemented a map-based user interface to query, select and edit interlinked web resources. qMAP (see **Figure 17**) allows the user to filter DBpedia [39] entries and related multimedia content like Flickr images [40], YouTube [44] videos or Last.fm music [46]. Thematically and geographically personalized knowledge views are possible. Knowledge gaming content can be also placed on the qMAP.

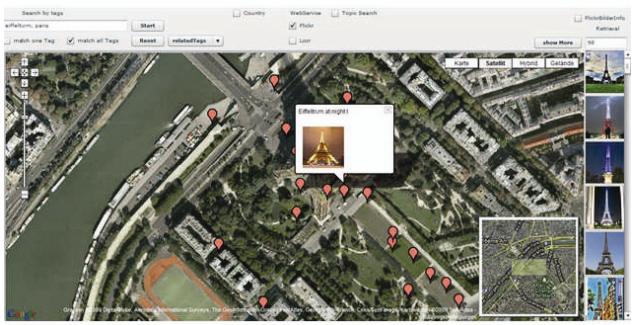


Figure 17. qMAP frontend

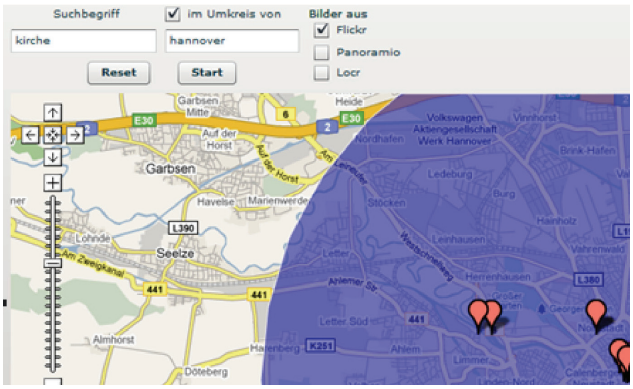



Figure 18. Search, filter and periphery interface of qMAP

Qualified Flickr images played first by qRANK are integrated into qMAP too (**Figure 17 and 18**). Figure 18 shows the periphery search and explore functionality of the qMAP. As shown in **Figure 19**, every user task and interaction is locked in qKAI's history protocol. Update, creation date or views of images are exemplary shown in Figure 19.

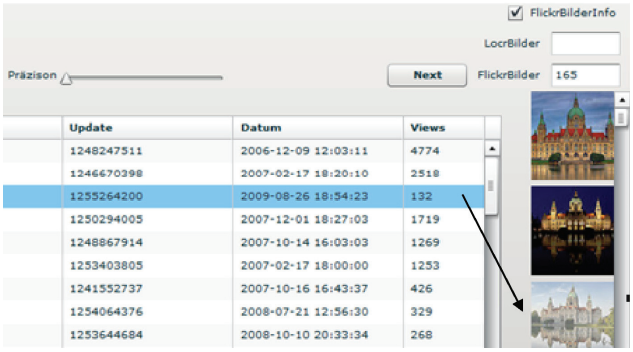



Figure 19. History and interaction protocol of Open Content for statistical analysis behind the qMAP interface.

The graphical interface of qMAP consists of three different states. So users can select with checkboxes individual functions or hide them. By default, a keyword-search in Flickr is set. The checkbox "*search by country*" the user can search for images within a certain radius and the checkbox "*topic search*" allows a search by topic. In order not to overload the map with markers, only a maximum of 100 images to each request is used. During the area search for Flickr images, the user has the additional option to set a radius (in km). The selected area is marked in blue on the map (see Figure 18).

### *B. qMATCH: An assignment quiz with Flickr content*

qMATCH [48] is a prototype of an image-term assignment gaming type. First, the user enters a term he likes to get images about. Then he gets presented randomized terms and images out of Flickr and he has to assign the right term to the right image via Drag & Drop assignment (see **Figure 20**).


Figure 20. qMATCH text-image assignment game

Here we need a service called wrong-answerizer to assign wrong, but not stupid answers. Wrong-answerizer is deployed in further gaming types. qMATCH is useful to enhance e.g., language skills, geographically, architectural or historical knowledge. If we use term-term assignment a lot of vocabulary out of various domains can be assessed: assigning English to German translations, assigning buildings to right historical epochs or assigning cities to the right countries. In **Figure 21**, the statistically protocol of a user and his interaction on Open Content like Flickr images is shown.

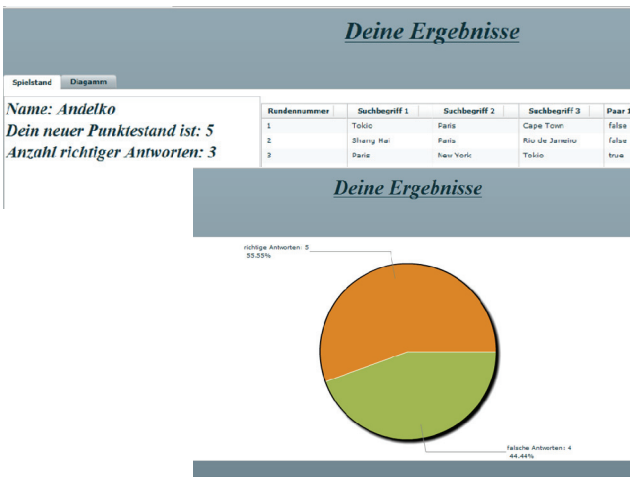



Figure 21. Knowledge game result in qMATCH with own correct answers and aggregated statistics.

## XII. CONCLUSION AND OUTLOOK

We have exemplified the role of information quality in web-based information and knowledge transfer with smart interaction. Beyond evaluating the state of the art, we adapted an existing assessment model to our purpose in qKAI and showed some examples for enhanced rating-based

interaction that is suitable to **qualify Open Content** stepwise in an incentive way. Incentive for user participation and interaction is implemented in qKAI as ontology-based, **global interaction rewarding system** for any kind of interaction (GIAR). All over, information quality enhancement is getting more and more important – especially regarding autonomous Web resources. Information quality can be utilized as a tool to derive personalization and user preferences in web-based information and knowledge systems, because it offers metrics to determine the fitness for use of autonomous, distributed resources.

The quality of the image search on the Web is a very topical subject of research. Many approaches and algorithms try to optimize the search. In this study, some possibilities are discussed and we implemented a **tag-based group ranking** method and a **game-based application** for the ranking of images. To show the effect of the procedure, images from Flickr were used. The focus of this contribution was the evaluation of user-generated metadata, which are derived from online communities, the so-called **tagging systems**. Especially for the search of images they are very important, because images, in contrast to other distributed content on the Web, do not contain metadata and are therefore difficult to find.

The simple form of tagging systems - free of any notation and relation of metadata generation - allows that content can be categorized by non experts. This, however, offers new challenges for Web search and data mining. The basic problem is to assess the relevance of the determined information. The advantage of the Semantic Web is that the information is in a machine-interpretable form because they were previously annotated semantically. It is different with metadata derived from the social annotation. Social annotation also called **collaborative tagging** arises when the common folk describe resources with keywords. In research, these are also known as **folksonomies**. To view the information from the folksonomies as useful advantage, they must be enriched with semantics. One possibility is to map them into lightweight ontologies. In this work, we discussed in detail how to combine folksonomies and tag ranking methods for images. The derived **keyword-oriented group search algorithms** and the **ranking game qRANK** are very promising, if the users are motivated to participate. Despite some weaknesses, tags are a useful addition to existing ontologies.

## REFERENCES

[1] Steinberg, M. and Brehm, J.: Towards enhanced user interaction to qualify Web resources for higher-layered applications, Proc. DigitalWorld 2010, International Conference on Mobile, Hybrid, and On-line Learning (IARIA's eLmL), Neth. Antilles, ISBN: 978-0-7695-3955-3, pp.105-110, 2010.

[2] Naumann, F.: Quality-Driven Query Answering for Integrated Information Systems, Lecture Notes in Computer Science, Vol. 2261, Springer, 2002.

[3] Steinberg, M. and Brehm, J.: Towards utilizing Open Data for interactive knowledge transfer, Proc. DigitalWorld 2009, International Conference on Mobile, Hybrid, and On-line Learning (IARIA's eLmL), IEEE Press, 2009, pp.61-66, doi:10.1109/eLmL.2009.13.

[4] Kruse, P.; Warnke, T.; Dittler, A.; and Gebel, T.: Wertewelt Medien, http://www.nextpractice.de/fileadmin/studien/medienstudie2007/Medienstudie_Nov2007.pdf, 2007.

[5] Open Knowledge Foundation, The Open Knowledge Definition, http://opendefinition.org/, last update: 2008, visited: 2011-01-12.

[6] Steinberg, M. and Brehm, J.: Social educational games based on Open Content, Proc. International Conference on Intelligent Networking and Collaborative Systems (INCoS), Spain, 2009.

[7] Juran, J.: The Quality Control Handbook. McGraw-Hill, New York, 3rd edition, 1974.

[8] Resource Description Framework, http://www.w3.org/RDF/, last update: 2009, visited: 2011-01-12.

[9] Wand, Y. and Wang. R.: Anchoring Data Quality Dimensions in Ontological Foundations, Communications of the ACM, 39(11):86–95, 1996.

[10] Wikipedia, www.wikipedia.en, last update: 2009, visited: 2011-01-12.

[11] Freebase, www.freebase.com, last update: 2009, visited: 2011-01-12.

[12] Amazon, www.amazon.com, last update: 2009, visited: 2011-01-12.

[13] Bizer, C.: Quality-Driven Information Filtering in the Context of Web-Based Information Systems, Dissertation, 2007.

[14] Heath, T. and Motta, E.: Revyu.com: A Reviewing and Rating Site for the Web of Data, Proc. ISWC 2007, International Semantic Web Conference, Lecture Notes in Computer Science 4825 Springer 2007, pp. 895-902.

[15] Mui, L.: Computational Models of Trust and Reputation: Agents, Evolutionary Games and Social Networks, Dissertation, 2003.

[16] Parker, M.; Moleshe, V.; De La Harpe, R.; and Wills, G.: An evaluation of Information quality frameworks for the World Wide Web, Cape Peninsula University of Technology, University of Southampton, http://de.scientificcommons.org/14463068, 2006.

[17] IMS/QTI, http://www.imsglobal.org/question/, IMS Global Learning Consortium, Inc., last update: 2008, visited: 2011-01-12.

[18] Jøsang, A.; Ismail, R.; and Boyd, C.: A Survey of Trust and Reputation Systems for Online Service Provision, 2006.

[19] Nov, O.: What motivates Wikipedians? Communications ACM, Vol. 50, Nr. 11, 2007.

[20] Aperture, http://aperture.sourceforge.net/, Aduna, DFKI, last update: 2008, visited: 2011-01-12.

[21] ISO 15836: 2003, Information and Documentation – The Dublin Core Metadata Element Set, International Organization for Standardization, 2003.

[22] Openlink Virtuoso, http://virtuoso.openlinksw.com/, last update: 2008, visited: 2011-01-12.

[23] OpenNLP, http://opennlp.sourceforge.net/, last update: 2008, visited: 2011-01-12.

[24] Sarioglu, O.: Design and implementation of a map-based frontend with geocoded knowledge units, master thesis, Leibniz Universität Hannover, System- and Computer Architecture, 2009.

[25] Abbasi, R. and Staab, S.: Richvsm: Enriched vector space models for folksonomies, Proc. HT 09, Hypertext and hypermedia, pp. 219–228, New York, NY, USA, 2009. ACM.

[26] Marlow, C.; Naaman, M.; Boyd, D.; and Davis, M.: Ht06, tagging paper, taxonomy, flickr, academic article, to read, Proc. HT 06, Hypertext and hypermedia, pp. 31–40, New York, NY, USA, 2006. ACM.

[27] Panke, T.; Gaiser, S.; and Hampel, B.: Good Tags – Bad Tags, Waxmann, 2008.

[28] Wu, L.; Yang, L.; Yu, N.; and Hua, X.: Learning to tag. In 18th International World Wide Web Conference, pp. 361–371, April 2009.

[29] Stock, W. and Peters, I.: Folksonomies in Wissensrepräsentation und Information Retrieval. In Information – Wissenschaft und Praxis 59 (2008) 2, pp. 77–90, 2008.

[30] Vogel, A.; Anderson, A.; and Raghunathan, K.: Tagez: Flickr tag recommendation. 2008.

[31] Hepp, M.; Coenen, T.; and Van Damme, C.: Quality metrics for tags of broad folksonomies, pp. 118–125, Proc. International Conference on Semantic Systems, Journal of Universal Computer Science, 2008.

[32] Hotho, A.; Jäschke, R.; Schmitz, C.; and Stumme, G.: Folkrank: A ranking algorithm for folksonomies, Proc. FGIR 2006, 2006.

[33] Abel, F.; Henze, N.; and Krause, D.: Context-aware ranking algorithms in folksonomies, Proc. Webist, pp. 167–174, 2009.

[34] Peinado, V.; Artiles, J.; Gonzalo, J.; Barker, E.; and Ostenero, F. L.: FlickLing: a multilingual search interface for Flickr, Working Notes for the CLEF 2008 Workshop, 2008.

[35] San Pedro, J., Siersdorfer, S.: Ranking and classifying attractiveness of photos in folksonomies, Proc. WWW '09, 18th international conference on World Wide Web, pp. 771–780, New York, NY, USA, 2009. ACM.

[36] Li, X.; Snoek C.; and Worring, M.: Learning tag relevance by neighbor voting for social image retrieval, Proc. MIR '08, 1st ACM international conference on Multimedia information retrieval, pp. 180–187, New York, NY, USA, 2008. ACM.

[37] Garg, N. and Weber, I.: Personalized tag suggestion for flickr, Proc. WWW '08, 17th international conference on World Wide Web, pp. 1063–1064, New York, NY, USA, 2008. ACM.

[38] Negoescu, R. and Gatica-Perez, D.: Analyzing flickr groups, Proc. CIVR '08: International conference on Content-based image and video retrieval, pp. 417–426, New York, NY, USA, 2008. ACM.

[39] DBpedia, www.dbpedia.org, last update: 2009, visited: 2011-01-12.

[40] Flickr, www.flickr.com, last update: 2010, visited: 2011-01-12.

[41] Abel, F.; Henze, N.; Krause, D.; and Kriesell, M.: On the effect of group structures on ranking strategies in folksonomies, Weaving Services and People on the World Wide Web, pp. 275–300, 2008.

[42] Abel, F.; Frank, M.; Henze, N.; Krause, D.; Plappert, D.; Siehnde, P.: Groupme! - where semantic web meets web 2.0, pp. 871–878, 2008.

[43] Abel, F.; Henze, N.; Krause, D.: Groupme! In WWW, pp. 1147–1148, 2008.

[44] YouTube, www.youtube.com, last update: 2009, visited: 2011-01-12.

[45] Del.icio.us, http://delicious.com, last update: 2010, visited: 2011-01-12.

[46] Last.fm, www.last.fm, last update: 2009, visited: 2011-01-12.

[47] Wang, G. and Hoiem, D.: Learning Image Similarity from Flickr Groups Using Stochastic Intersection Kernel Machines, Proc. ICCV 2009.

[48] Jovancevic, A.: Analysis and extension of interaction with Open Content in the Social Semantic Web, master thesis, Leibniz Universität Hannover, System- and Computer Architecture, 2009.

[49] Cattuto, C.; Benz, D.; Hotho, A.; Stumme, G.: Semantic analysis of tag similarity measures in collaborative tagging systems; Proc. of the 3rd Workshop on Ontology Learning and Population (OLP3), July 2008.

[50] Specia, L. and Motta, E.: Integrating folksonomies with the semantic web, pp. 624–639, 2007.

[51] Sigurbjörnsson, B. and van Zwol, R.: Flickr tag recommendation based on collective knowledge, Proc. WWW '08, 17th international conference on World Wide Web, pages 327–336, New York, NY, USA, 2008, ACM.

# Bag Relational Algebra with Grouping and Aggregation over C-Tables with Linear Conditions

Lubomir Stanchev
*Computer Science Department*
*Indiana University - Purdue University Fort Wayne*
*Fort Wayne, IN, USA*
*stanchel@ipfw.edu*

***Abstract*—We introduce bag relational algebra with grouping and aggregation over a particular representation of incomplete information called c-tables, which was first introduced by Grahne in 1984. In order for this algebra to be closed and "well-defined", we adopt the closed world assumption as described by Reiter in 1978 and extend the tuple and table conditions to linear ones. We explore the problem of rewriting and simplifying this novel type of c-tables, show how to perform equivalence test for c-tables, and argue why it is difficult to create a canonical form for c-tables. We present certain answer semantics for a full-blown relational algebra with grouping and aggregation and accordingly present algorithms for executing the different relational algebra operators over our representation of incomplete information. The algorithms run in polynomial time relative to the size of the precise information, which makes them a candidate for implementation as part of a DBMS engine that supports storage and retrieval of incomplete information.**

***Keywords-incomplete information*; *c-tables*; *relational model*; *null values*; *bag semantics***

## I. INTRODUCTION

This paper extends a conference paper on the topic of querying incomplete information ([1]). We have added theoretical results on simplifying and checking the equivalence of c-tables and discussion on the existence of a canonical form for c-tables. We have also expanded the description of all algorithms that implement non-trivial relational algebra operators, such as monus, grouping, and aggregation, and added detailed proofs on the correctness and time complexity to all algorithms.

Many times, when information is entered into databases, the values for some of the fields are left empty for various reasons. In some cases, partial information about the blank fields is available. However, existing relational database technology does not allow for such information to be processed. Imielinski and Lipski in [2] were among the first to propose richer semantics for null values that allows for incomplete information to be processed. However, their model was based on set semantics. Later on, Libkin and Wong published a paper on querying incomplete information in databases with multisets ([3]), but included only a limited set of operators that excluded grouping and aggregation. Other papers that tackle the problems of storing and querying incomplete information include [4], [5], [6], [7], [8]. However, they all fail to explore grouping and aggregation over bag semantics.

In this paper, we fill a gap in published research in the area of storing and querying incomplete information. More precisely, we show how bag relational algebra with grouping and aggregation can be applied over incomplete information represented as a particular variation of *c-tables*. A c-table consists of a of *c-tuples* and a *global condition*, where every c-tuple contains a regular tuple that may include variables for some of its fields plus a local condition (See Table I for an example). The semantics of a c-table is determined by the set of relational tables that it represents, where each representation is derived from a valuation for the variables in the c-table. In order for the relational algebra over c-tables to be closed and *well defined*, we define the semantics of a c-table to be over the *closed world assumption*, as defined in [9], and we extend local and global conditions to be linear. We will refer to such c-tables as *linear c-tables*, where the exact semantics will be presented in Section II-A.

C-tables were first introduced by Grahne in [10] to have local and global conditions that did not contain the "+" operator and the ">" relation. Later on, Grahne added the ">" relation in [4]. However, we are not aware of any published research that allows for the "+" operator to be part of the local or global condition of a c-table. On the other hand, introducing the "+" operator is required in order for relational algebra with aggregation over c-tables to be closed.

Note that several different linear c-tables may have the same semantics, that is, have the same set of representations. This is the reason why it is desirable to be able to check for equivalence between linear c-tables and be able to normalize linear c-tables. For example, when we store or visualize a linear c-table, we would want to use a compact and easy to understand representation. In the paper we present a novel procedure for simplifying linear c-tables that runs in polynomial time relative to the size of the precise information. We show why it is difficult to construct a canonical form for linear c-tables and solve the problem of comparing linear

c-tables for equality.

The main contributions of the paper are the algorithm for simplifying linear c-tables, the algorithm for comparting two c-tables for equality, and the algorithms for performing the different relational operators over linear c-tables. While the implementation of the operator projection, selection, and inner join are similar to the case of set semantics (see [2]), the algorithms for monus, duplicate elimination, grouping, and aggregation are non-trivial and novel.

### A. Motivation

Real world requirements have shown the importance of storing and querying incomplete information. However, contemporary database management systems (DBMSs) provide only limited support (that is, only null values). Part of the reason is the lack of research in the area. While the problem of storing incomplete information is somewhat solved, querying incomplete information remains an open research challenge. This paper makes a significant step towards solving the later problem.

The main hurdle towards the implementation of a DBMS that can processes rich incomplete information is the intrinsic high cost of managing such information. However, note that the algorithms that we present for performing the various relational algebra operators are non-polynomial relative only to the size of the incomplete information. Taking into account the ever-increasing speed of computational resources, we believe that incorporating tools that store and query incomplete information within commercial database engines is feasible and practical. This work can play a key part in such an endeavor. For example, since the code for executing bag relational algebra operators is an important part of the kernel of a SQL engine, our algorithms can be used to implement a SQL engine that can query incomplete information stored as linear c-tables.

In what follows, in Section II we define a representation of incomplete information in terms of linear c-tables. In Section III we describe how linear c-tables can be simplified and compared for equality and explore the problem of existence of canonical form for c-tables. In Section IV we define bag relational algebra operators over linear c-tables and present example algorithms for their implementation. In Section V the problems of grouping and aggregation over linear c-tables are explored. Section VI provides a summary or the presented work and addresses areas for future research.

## II. C-Tables with Linear Conditions

The problem of representing incomplete information in the relational model is almost as old as the relational model itself ([11], [12], [13], [14], [15]). When a null value appears in a relational table, its value can be interpreted as no information available, only partial information available, value not applicable, and so on. Most of the research on null values has concentrated on the first two meanings. Known representations of relational tables adapting these meanings for nulls include Codd tables, naïve tables, Horn tables and c-tables. Codd tables are relational tables, where the values of some the fields can be null. Naïve tables are an extension of Codd tables, where each null is given a label and nulls having the same label represent the same unknown value. C-tables are naïve tables with a local condition associated which each c-tuple and a single global condition associated with each c-table. A c-tuple in a c-table is part of the representation of the c-table under some valuation when the local condition of the c-tuples and global condition of the c-table are both true. Horn tables are a special kind of c-tables in which the local and global conditions are restricted to Horn clauses.

Grahne, in [4], considered Boolean conditions over the system $\langle R, \{>, =\}\rangle$ (i.e., Boolean expressions with variables and constants defined over the set $R$ extended with "$>$" and "$=$"). To the best of our knowledge, except for [1], this is the most expressive system for expressing c-table conditions in published research.

In this paper we explore c-tables with conditions over the system $\langle \mathbb{R},\ \{>,\ =,\ +\}\rangle \cup \langle \mathbf{S}, \{=, \neq\}\rangle$, where $\mathbb{R}$ is used to denote the set of real numbers and $\mathbf{S}$ is the set of strings over some finite alphabet. While the "$+$" operator is introduced in order to make the algebra closed relative to aggregation, the system over strings is introduced in order to extend the expressive power of c-tables. Note that we do not explore conditions over $\langle \mathcal{Z},\ \{>,\ =,\ +\}\rangle$, where $\mathcal{Z}$ is the set of integers, or over $\langle \mathbb{R},\ \{>,\ =,\ *,\ +\}\rangle$. The reason is that, although these systems are more expressive, reasoning with them is much harder. For example, Fischer and Rabin have shown that the time complexity of deciding whether a formula over the first system is satisfiable is super exponential ([16]). Similarly, the time complexity of the fastest known algorithm for solving the same problem for the second system, which is presented in [17], is higher than exponential.

### A. Definitions

We next present the syntax and semantics of a linear c-table.

*Definition 1 (syntax of linear c-table):* A *linear c-table* $T$ as a finite and unordered bag of *linear c-tuples* and a *global condition*[1]. A linear c-tuple with attributes $\{A_i\}_{i=1}^{a}$ is the sequence of mappings from $A_i$ to $D(A_i) \cup V_i$ plus a local condition, where $i$ ranges from 1 to $a$, $D(A_i)$ denote the domain of $A_i$ and $V_i$ is used to represent a possibly infinite but countable set of variables over $D(A_i)$. The local and global conditions can range over the system $\langle \mathbb{R}, \{>, =, +\}\rangle \cup \langle \mathbf{S}\ \{=, \neq\}\rangle$.

[1]In order to keep the notation simple, we do not use special syntax for c-tuples and c-tables, where it will be clear from the context when we are referring a c-table (c-tuple) and when to a relational table (tuple).

| *name* | *school* | *condition* |
|---|---|---|
| John | $y$ | $x = 1$ |
| Mark | $y$ | $x \neq 1$ |
| $q$ | $z$ | TRUE |

g.c. $(q \neq$"Mark"$) \wedge (q \neq$"John"$) \wedge (z \neq y)$

Table I
AN EXAMPLE LINEAR C-TABLE

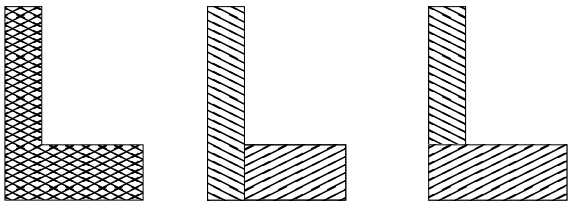


Figure 1. Three different ways to represent the same two-dimensional point set as union of polyhedra

Table I shows an example of a linear c-table. We will refer to the part of a linear c-table where the data is stored as the *main part* and to the remaining parts as the *local condition part* and the *global condition part*, respectively. In Table I, $x$, $y$, $z$ and $q$ are used to represent variables. Since our model is limited only to the domains of real numbers and strings, the domain of a variable that does not appear in the main part of a linear c-table can be inferred from the context in which it appears. Fore example, we can use the local condition $x = 1$ to deduce that the domain of $x$ is the set of real numbers.

Table I expresses the information that either there are no students or there are two students that study in different schools and the name of one of them is "John" or "Mark" and the name of the other one is neither "John" nor "Mark". Note that in this example and throughout the paper we will be using the closed world assumption. The assumption states that the database contains all existing individuals. In our example, we have used this assumption to conclude that there are at most two students in the database.

In order to formally define the semantics of a c-table, Imielinski and Lipski introduce a function called *rep* that maps a c-table $T$ to a possibly infinite set of relational tables ([2]). Intuitively, the meaning of the *rep* function is that given a c-table $T$, the function returns all relational tables that $T$ represents under different valuations. In [2], this function is defined relative to the *open world assumption*. We define it relative to the closed world assumption. In the definition that follows, $main$, $lc$ and $gc$ are used to denote the main part, the local condition part, and the global condition part of a linear c-table, respectively. The symbol $\varepsilon$ is used to denote the empty set.

*Definition 2 (semantics of a linear c-table):* A linear c-table $T$ represents the set of relational tables that are defined by the following equation.

$$rep(T) = \{T' | \exists v, \text{ such that } \ v(T) = T'\} \qquad (1)$$

In the definition, $v$ is a mapping that maps the variables in $T$ to constants in the corresponding domains and is generalized to linear c-tuples as follows.

$$v(t) = \begin{cases} v(main(t)) & : \ v(lc(t)) \wedge v(gc(T)) \\ \varepsilon & : \ otherwise \end{cases} \qquad (2)$$

The value of $v(main(t))$ is calculated by substituting the variables in the main part of $t$ with the values to which $v$ maps them. The mapping $v$ is further extended to linear c-tables as shown in Equation 3, where $\{t_i\}_{i=1}^{k}$ are the linear c-tuples in $T$.

$$v(T) = \{| \ v(t_i) | i \in [1, k] \wedge v(t_i) \neq \varepsilon \ |\} \qquad (3)$$

In the above definition, we have used the common notation $\{| \cdot |\}$ to denote a bag of elements. While it is possible to define an ordering on the linear c-tuples inside a linear c-table, we leave this topic as area for future research.

Definition 2 is novel and differs from the definitions presented in [2], [4]. Unlike these papers, we define duplicate semantics for c-tables and use the closed world assumption.

From now on, when the distinction is clear from the context, we will refer to linear c-tuples simply as c-tuples and to linear c-tables simply as c-tables.

## III. SIMPLIFYING LINEAR C-TABLES

An important part of simplifying a linear c-table is simplifying the local conditions and the global condition, which are both expressed as linear conditions, and checking for their satisfiability. Details on how to simplify a linear condition and how to check if it is satisfiable under at least one valuation are presented next.

### A. Linear Condition Simplification and Satisfiability Check

A *linear condition* is a Boolean expression and, as such, can be expressed as a disjunction of *positive conjunctions*. A positive conjunction is a conjunction of positive atomic linear conditions, where the later has the form $\bar{a} \cdot \bar{x} = \bar{b}$ or $\bar{a} \cdot \bar{x} < \bar{b}$ ($\bar{x}$ is a variable vector and $\bar{a}$ and $\bar{b}$ are vector constants). An *atomic linear condition* includes in addition negative conditions of the form $\bar{a} \cdot \bar{x} \neq \bar{b}$. An intuitive representation of a positive conjunction is a multi-dimensional polyhedron, which defines a semilinear set. Therefore, a linear condition can be interpreted as a set of disjoint polyhedra. Note however that, as shown if Figure 1, such a representation is not unique.

Let us first consider the algorithm that was proposed in [18] for normalizing conjunctions of linear equalities and inequalities. More precisely, the paper represents a conjunction of atomic linear conditions by the system $A\bar{x} \leq \bar{b}$, $E\bar{x} = \bar{d}$, $\neg(\bar{c}_i\bar{x} = \bar{f}_i)$, where $A$ and $E$ are matrices with constants, $\bar{b}$, $\bar{d}$, $\bar{c}_i$ and $\bar{f}_i$ are vectors of constants and $\bar{x}$ is a variable vector. The normalization algorithm

runs in polynomial time and relies on calls to a module that solves linear programs. We will refer to this algorithm as *normalize*.The algorithm has the added advantage that it recognizes sets of unsatisfiable atomic conditions and reports them as such by returning the empty set. Part of the algorithm deals with the elimination of redundant conditions, which is an extension of the research that is published in [19]. The pivot theorem from [18] follows.

*Theorem 1:* If two sets of atomic conditions over $\langle \mathbb{R}, \{+, >, =\} \rangle$ define the same point set, where $\mathbb{R}$ is the set of real numbers, then their canonical forms will have identical set of equality conditions, the same inequality conditions up to multiplication by a positive scalar, and the same set of negative conditions.

**Algorithm 1** *simplify*($C$)

1: $c_1 \vee c_2 \vee ... \vee c_n \leftarrow C$, where $\{c_i\}_{i=1}^{n}$ are positive conjunctions.
2: *result* $\leftarrow$ *break_up*$(c_1, \ldots, c_n)$
3: **if** *result* $= \emptyset$ **then**
4: **return false**
5: **end if**
6: **return** $g_1 \vee \cdots \vee g_m$, where $\{g_i\}_{i=1}^{m}$ are the conjunctions in *result*.

**Algorithm 2** *break_up*$(c_1, \ldots, c_n)$

1: *result* $\leftarrow \{$*normalize*$(c_1)\}$
2: **for** $i \leftarrow 2$ to $n$ **do**
3: **for** $g \in$ *result* **do**
4: *result* $\leftarrow$ *result* $\cup \{$*normalize*$(g \wedge c_i)\}$
5: *result* $\leftarrow$ *result* $\cup \{$*normalize*$(g \wedge \neg c_i)\}$
6: *result* $\leftarrow$ *result* $\cup \{$*normalize*$(\neg g \wedge c_i)\}$
7: *result* $\leftarrow$ *result* $- \{g\}$
8: **end for**
9: **end for**
10: **for** $g \in$ *result* **do**
11: **if** $g =$ **false then**
12: *result* $\leftarrow$ *result* $- \{g\}$
13: **end if**
14: **end for**
15: **return** *result*.

The pseudo-code for simplifying a linear condition $C$ is presented in Algorithm 1. The algorithm first breaks $C$ into a disjunction of positive conjunctions. Next, the algorithm divides the conjunctions so that they do not overlap. As a final step, the algorithm normalizes the conjunctions that are computed using the normalization algorithm from [18]. The following theorem address the correctness of the algorithm.

*Theorem 2:* Algorithm 1 is correct, that is, $C =$ *simplify*$(C)$ for any linear condition $C$.

*Proof*: Line 1 of Algorithm 1 breaks $C$ into disjunctive normal form and therefore does not change the value of $C$. Algorithm 2 breaks up conjunctions so that they do not overlap. The conjunction of the expressions $g \wedge c_i$, $g \wedge \neg c_i$, and $\neg g \wedge c_i$ is equal to $g \vee c_i$. Therefore, Lines 4-7 of the Algorithm 2 remove $g$ from the set of conjunctions stored in *r*esult and add $g \vee c_i$. Therefore, the net effect of the lines is to add $c_i$ to *result*. Therefore, after Lines 1-9 of Algorithm 2 the conjunctions $\{c_i\}_{i=1}^{n}$ are added to *result*. Lines 10-14 of Algorithm 2 remove `false` conjunctions from *result*. If after this process *result* is empty, then $C$ is not satisfiable and `false` is returned correctly at Lines 4 of Algorithm 1. Line 6 of Algorithm 1 returns the computed disjoint conjunctions. ■

*Theorem 3:* Algorithm 1 runs in $O(m^c \cdot 3^n)$ time, where $m$ is the length of the linear condition $C$, $n$ is the number of conjunctions in the disjunctive normal form of $C$ (i.e., $n \leq (\sqrt{2})^m$), and $c$ is a constant.

*Proof*: In order to verify the running time of the algorithm, note that Line 1 takes $O(m \cdot n)$ time. Line 1 of Algorithm 2 makes a call to the normalization procedure from [18], which runs in $O(m^c)$ time (the length of each conjunction is smaller then the length of $C$). Lines 2-9 of Algorithm 2 make at most $\frac{3^n-1}{2} - 1$ calls to the procedure from [18]. The reason is that during the $k^{\text{th}}$ iteration of the outer `for`-loop there can be as many as $3^{(k-2)}$ conjunctions in *result* and therefore as much as $3^{(k-1)}$ calls to the normalization procedure from [18]. Lines 10-14 of Algorithm 2 take less time to execute than Lines 2-9 of Algorithm 2 and therefore do not contribute to the complexity. Therefore, $\sum_{k=2}^{n} 3^{k-1} = \frac{3^n-1}{2} - 1$ is an upper bound on the number of calls to the normalization procedure and each call takes $O(m^c)$ time. ■

Algorithm 1 can be used to test the satisfiability of a linear condition. However, the part of the algorithm that removes the overlapping part of the conjunctions will no longer be needed. The modified pseudo-code is shown in Algorithm 3. We will refer to the simplified algorithm as *fast_simplify*. Alternative methods for testing for linear condition satisfiability are described in [20], [21]. The full-blown algorithm is only useful when we want to eliminate including the same point set multiple times by breaking up a linear condition into disjoint polyhedra.

**Algorithm 3** *fast_simplify*($C$)

1: $c_1 \vee c_2 \vee ... \vee c_n \leftarrow C$, where $\{c_i\}_{i=1}^{n}$ are positive conjunctions.
2: **for** $i \leftarrow 1$ to $n$ **do**
3: $c_i \leftarrow$ *normalize*$(c_i)$
4: **end for**
5: **if** $c_i =$ **false** for $i = 1$ to $n$ **then**
6: **return false**
7: **end if**
8: **return** $c_1 \vee \cdots \vee c_n$

*Theorem 4:* Algorithm 3 is correct, that is $fast\_simplify(C) = C$. Moreover, $fast\_simplify(C)$ returns `false` exactly when $C$ is not satisfiable.

*Proof*: The algorithm breaks $C$ into disjunctive normal form and normalizes each conjunction. This will not affect the value of $C$. Note that when $C$ is not satisfiable each conjunction $c_i$ will be evaluated as `false` (see [18] for a formal proof) and therefore the method will return `false`. ■

*Theorem 5:* The running time of Algorithm 3 is $O(n \cdot m^c) = O((\sqrt{2})^m \cdot m^c)$, where $m$ is the length of the linear condition $C$, $n$ is the number of conjunctions in the disjunctive normal form of $C$ (i.e., $n \leq (\sqrt{2})^m$), and $c$ is a constant.

*Proof*: Line 3 is executed $n$ times and the complexity of the *normalize* method is $O(m^c)$. ■

The algorithm can be applied not only to conditions over the system $\langle \mathbb{R}, \{>, =, +\}\rangle$, but also to conditions over the system $\langle \mathbb{R}, \{>, =, +\}\rangle \cup \langle \mathbf{S}, \{=, \neq\}\rangle$. To do so, substitute each atomic conditions of the form $x \neq c$, where $x$ is a string variable and $c$ is a string constant with $x = c_1 \vee x = c_2 \vee \cdots \vee x = c_r \vee x = c_{r+1}$, where $c_{r+1}$ is a newly introduced string constant and $\{c_i\}_{i=1}^{r}$ are the existing string constants excluding $c$. In other words, we pin the value of $x$ to be equal to one of the existing constants (excluding $c$) or to a new constant, which is equivalent to stating that $x \neq c$.

Similarly, substitute each atomic condition of the form $x \neq y$, where $x$ and $y$ are string variables with $\bigvee_{i,j=1,r+2}^{i \neq j} (x = c_i \wedge y = c_j)$, where $c_{r+1}$ and $c_{r+2}$ are newly introduced constants and $\{c_i\}_{i=1}^{r}$ are the existing string constants. In other words, we add the restriction on the variables $x$ and $y$ that they are equal to distinct constants, which implies $x \neq y$.

Alternatively, Line 1 of the algorithm can be modified to require the breaking of $C$ into not necessarily positive conjunctions. This modification allows the direct application of the normalization algorithm to a linear condition containing strings because the algorithm from [18] handles inequality conditions in addition to equality and weak-inequality (i.e., greater than and less than) conditions.

### B. C-Table Simplification

Note that there may be different c-tables representing the same set of bag relational tables, that is, it may be the case that $T_1 \neq T_2$ but $rep(T_1) = rep(T_2)$. The following definition formally defines the concept of c-table equivalence.

*Definition 3 (c-table equivalence):* If $rep(T_1) = rep(T_2)$, then we will say that $T_1$ and $T_2$ are equivalent and write $T_1 \approx T_2$.

Algorithm 4 shows how to simplify a c-table. The algorithm relies on the notion of c-tuple unification, which is defined next.

| *A* | *B* | *condition* |
|---|---|---|
| 1 | 2 | $x = 1$ |
| $z$ | 2 | $x = 2$ |
| $p$ | $w$ | $x = t$ |

g.c.: $t \neq 1 \wedge t \neq 2$

| *A* | *B* | *condition* |
|---|---|---|
| $a$ | $b$ | $((a = 1) \wedge (b = 2) \wedge (x = 1)) \vee$<br>$((a = z) \wedge (b = 2) \wedge (x = 2)) \vee$<br>$((a = p) \wedge (b = w) \wedge (x = t))$ |

g.c.: $t \neq 1 \wedge t \neq 2$

Table II
A C-TABLE AND THE RESULT OF APPLYING STEPS 1 AND 2 OF THE C-TABLE SIMPLIFICATION ALGORITHM

*Definition 4 (c-tuple unification):* The c-tuples $t_1$ and $t_2$ of the c-table $T$ are unifiable exactly when the formula $lc(t_1) \wedge lc(t_2) \wedge gc(T)$ is not satisfiable. We will denote this check as *unifiable*, that is $unifiable(t_1, t_2) = \neg(lc(t_1) \wedge lc(t_2) \wedge gc(T))$.

**Algorithm 4** *simplify*($T$)

1: **for** $t \in T$ **do**
2: **if** *fast_simplify*($lc(t) \wedge gc(t)$) = **false then**
3: remove $t$ from $T$
4: **end if**
5: **end for**
6: **while** $\exists\{t_1, t_2\}$ s.t. *unifiable*$(t_1, t_2)$ **do**
7: remove $t_1$ and $t_2$ from $T$
8: create c-tuple $t$ with main part $\bar{X} = x_1, x_2, ..., x_n$, where $n$ is the arity of $T$ and $\{x_i\}_{i=1}^{n}$ are newly introduced variables.
9: add to $t$ the local condition $(\bar{X} = main(t_1) \wedge lc(t_1)) \vee (\bar{X} = main(t_2) \wedge lc(t_2))$
10: add $t$ to $T$
11: **end while**
12: **for** $t \in T$ **do**
13: $lc(t) \leftarrow$ *simplify*($lc(t) \wedge gc(T)$)
14: **if** $lc(t) =$ **false then**
15: remove $t$ from $T$
16: **else**
17: **while** $main(t)$ contains the variable $x$ for an attribute and *fast_simplify*($lc(t) \Rightarrow (x = c)$) = **true do**
18: replace $x$ with constant $c$ in $main(t)$
19: **end while**
20: **end if**
21: **end for**
22: $gc(T) \leftarrow$ **true**
23: **return** $T$

The intuition behind the definition is that if two c-tuples have local conditions that cannot both hold under any valuation, then at most one of the c-tuples could be present in any representation of the c-table and therefore the two

c-tuples can be merged into a single c-tuple.

Table II shows the result of applying the first eleven lines of the algorithm. The following theorem captures the correctness of Algorithm 4.

*Theorem 6:* Algorithm 4 is correct, that is, $simplify(T) \approx T$ for any c-table $T$.

*Proof*: Lines 1-5 of the algorithm remove c-tuples that are not part of any representation. Therefore, they will not have an effect on *rep*$(T)$. Lines 6-11 of the algorithm unify c-tuples that can be unified and thus reducing the size of the c-table without changing its representations. The reason is that two c-tuples that have incompatible local conditions cannot both appear in any representation. Lines 12-22 of of the algorithm move the global condition to the local conditions and simplify the resulting local conditions. Again, this will not affect the set of representations for the table. ■

The next theorem describes the time complexity of Algorithm 4.

*Theorem 7:* Algorithm 4 runs in $O(d^3 \cdot (\sqrt{2})^{d \cdot m} \cdot (d \cdot m)^c + (m \cdot d)^c \cdot 3^{(\sqrt{2})^{m \cdot d}} \cdot n)$ time, where $n$ is the number of c-tuples, $m$ is the greater of the size of the longest c-tuple and the size of the global condition, $d$ is the number of c-tuples with non-trivial local conditions (i.e., local conditions that are different than `true`), and $c$ is a constant.

*Proof*: Lines 1-5 of the algorithm takes $O((\sqrt{2})^m \cdot m^c \cdot d)$ time. The reason is that algorithm *fast_simplify*, which takes $O((\sqrt{2})^m \cdot m^c)$ time, needs to be applied to $d$ local conditions.

Lines 6-11 will take $O(d^3 \cdot (\sqrt{2})^{d \cdot m} \cdot (d \cdot m)^c)$ time. The reason is that $\binom{d}{2} + \binom{d-1}{2} + \cdots + \binom{2}{2} = O(d^3)$ iterations of the while loops can be performed and each iteration can take as much as $O((\sqrt{2})^{d \cdot m} \cdot (d \cdot m)^c)$ time because we use the *fast_simplify* algorithm on expressions as long as $d \cdot m$ when checking if two c-tuples are unifiable.

Line 13, which has the highest time complexity in the loop defined by Lines 12-21, can be applied on a local condition as big as $m \cdot d$ and therefore takes $O((m \cdot d)^c \cdot 3^{(\sqrt{2})^{m \cdot d}})$ time. The line can be applied at most $n$ times.

Line 22 can be be executed in constant time. ■

Note that Algorithm 4 can be improved by applying dynamic program or iterative dynamic programming techniques ([22]). For example, we can buffer existing results and use them in performing new calculations. This approach will save time because most of the presented algorithms produce c-tuples with local conditions that have subexpressions in common.

Note as well that Algorithm 4 does not produce a canonical form for c-tables. In order to understand why it is challenging to create a canonical form for c-tables, consider the first c-table from Table III. As the table shows, there are two different ways to apply Algorithm 4. Since the algorithm is non-deterministic, applying the algorithm differently yields different results. Therefore, the purpose of Algorithm 4 is not to find a canonical form for c-tables but to simplify a c-table. Since, as Table III suggest, unifying two c-tuples can prevent us from unifying one of the c-tuple with a third c-tuple, the problem of finding a canonical form for c-tuples is intrinsically hard.

| *A* | *condition* |
|---|---|
| 1 | $x < 2$ |
| 1 | $3 < x < 5$ |
| 1 | $4 < x < 6$ |

| *A* | *condition* |
|---|---|
| 1 | $x < 2 \vee 3 < x < 5$ |
| 1 | $4 < x < 6$ |

| *A* | *condition* |
|---|---|
| 1 | $x < 2 \vee 4 < x < 6$ |
| 1 | $3 < x < 5$ |

Table III
AN EXAMPLE C-TABLE SIMPLIFICATION

### C. Checking for C-Table Equality

As we have seen so far, c-tables are different from relational tables because they allow multiple ways to represent the same information. Therefore, checking for c-table equality is not trivial. The following theorem describes one possible way to do so.

*Theorem 8:* Two c-tables $T_1$ and $T_2$ represent the same set of tables exactly when *simplify*$(T_1 \dot{-} T_2) = \emptyset$ and *simplify*$(T_2 \dot{-} T_1) = \emptyset$, where "$\dot{-}$" is the monus operation that is introduced in the next section.

*Proof*: $\Rightarrow$ Let $T_1$ and $T_2$ represent the same set of tables. Then *simplify*$(T_1 \dot{-} T_2) \approx \emptyset$ and *simplify*$(T_2 \dot{-} T_1) \approx \emptyset$. However, note that if a c-table represents the empty set, then all local conditions will be unsatisfiable after Line 13 of Algorithm 4 and Lines 14-15 will remove all c-tuples from the c-table and make it empty. Therefore, it will be the case that *simplify*$(T_1 \dot{-} T_2) = \emptyset$ and *simplify*$(T_2 \dot{-} T_1) = \emptyset$.

$\Leftarrow$ Let *simplify*$(T_1 \dot{-} T_2) = \emptyset$ and *simplify*$(T_2 \dot{-} T_1) = \emptyset$. Then $rep(T_1 \dot{-} T_2) = \emptyset$ and $rep(T_2 \dot{-} T_1) = \emptyset$ and therefore it must be the case that $T_1$ and $T_2$ represent the same set of relational tables. ■

## IV. BAG RELATION ALGEBRA FOR C-TABLES

So far, we have defined the syntax and semantics of c-tables and presented an algorithm for their simplification. Next, we will describe how relational algebra[2] can be extended to handle c-tables. Specifically, since we are using the closed world assumption, we are able to develop a *sound* and *complete* extension of relational algebra that is *closed*. The definition of three terms follows.

*Definition 5 (closed relational algebra):* A relational algebra is *closed* exactly when the result of applying any operator $q$ with arity $n$ of the relational algebra to the c-tables $\{T\}_{i=1}^{n}$ produces a c-table, that is, $q(T_1, \ldots, T_n)$ is always a c-table.

[2]Our choice of relational algebra is arbitrary, that is, any language with the expressive power of relational algebra, such as relational calculus, can be used instead.

*Definition 6 (sound relational algebra):* A relational algebra is *sound* exactly when only correct answers appear in the result of $q(T_1, \ldots, T_n)$ or formally $rep(q(T_1, \ldots, T_n)) \subseteq q(rep(T_1, \ldots, T_n))$ for any c-tables $\{T_i\}_{i=1}^n$ and operator $q$ with arity $n$.

Note that throughout the paper we use $q(rep(T_1, \ldots, T_n))$ to denote the result of applying $q$ to each table in the set $rep(T_1, \ldots, T_n)$.

*Definition 7 (complete relational algebra):* A relational algebra is *complete* exactly when all correct answers appear in the result of $q(T_1, \ldots, T_n)$ or formally $q(rep(T_1, \ldots, T_n)) \subseteq rep(q(T_1, \ldots, T_n))$ for any c-tables $\{T_i\}_{i=1}^n$ and operator $q$ with arity $n$.

A relational algebra operator is *well defined* exactly when it is closed, sound, and complete. In this section we define the semantics of *projection*, *selection*, *inner join*, *union*, *monus*, and *duplicate elimination* over c-tables with bag semantics and show that all operators are well defined. The grouping and aggregation operations are discussed in the next section.

### A. Projection

*Definition 8 (syntax and semantics of projection):* If $T$ is a c-table with attributes $\bar{A}$, then we denote the projection of the attributes $\bar{A}'$ over this c-table as $\pi_{\bar{A}'}(T)$. The pseudo-code for performing the projection operator is shown in Algorithm 5.

The c-table $\pi_{\bar{A}'}(T)$ is constructed from the c-table $T$ by removing all columns in $\bar{A} - \bar{A}'$ and leaving the same local and global conditions.

**Algorithm 5** $\pi_{\bar{A}'}(T)$

```
for t ∈ T do
    remove attributes outside the set A from t
end for
return T
```

Note that the above definition defines duplicate-preserving projection. The duplicate-eliminating projection, which is more common in the relational model, can be constructed by applying the duplicate-elimination operator to the result of applying the duplicate-preserving projection. Algorithm 5 does not remove conditions that include variables associated with removed attributes because these conditions are still relevant. For example, even if an attribute that contains the variable $x$ is removed from Table I, the variable $x$ should not be removed from the local conditions because it stores the information that only one of the first two c-tuples can appear in any representation.

*Theorem 9:* The projection operator is well defined.

*Proof*: We need to show that $rep(\pi_{\bar{A}}(T)) = \pi_{\bar{A}}(rep(T))$.
$\Rightarrow$ Let $T_1 \in rep(\pi_{\bar{A}}(T))$, where $T_1$ is a relational table. Then there exists a valuation $v$ such that $T_1 = v(\pi_{\bar{A}}(T))$. Let $T_2$ be a relational table that extends $T_1$ with arbitrary values for the attributes outside the set $\bar{A}$. Then the equation $T_1 = \pi_{\bar{A}}(T_2)$ will hold. Let $v'$ be the valuation $v$ extended so that $T_2 = v'(T)$. Then $T_2 \in rep(T)$ and therefore $T_1 \in \pi_{\bar{A}}(rep(T))$.

$\Leftarrow$ Let $T_1 \in \pi_{\bar{A}}(rep(T))$. Then there exists valuation $v$ such that $T_1 = \pi_{\bar{A}}(v(T))$. Let $T_2$ be a relational table that extends $T_1$ with arbitrary values for the attributes outside the set $\bar{A}$. Then the equation $T_1 = \pi_{\bar{A}}(T_2)$ will hold. Note that $T_1 = v'(\pi_{\bar{A}}(T))$ where $v'$ is a valuation that extends $v$ to the attributes outside the set $\bar{A}$ according to the values of the attributes in $T_2$. Therefore $T_1 \in rep(\pi_{\bar{A}}(T))$. ■

| $A$ | $B$ | *condition* |
|---|---|---|
| 2 | $x$ | $x \neq 3$ |
| 2 | 4 | TRUE |

g.c. $x \neq 2$

| $B$ | $C$ | *condition* |
|---|---|---|
| 4 | 1 | TRUE |
| 2 | $z$ | $z > 3$ |

g.c. TRUE

Table IV
EXAMPLE $R_1$ AND $R_2$ C-TABLES

| $B$ | *condition* |
|---|---|
| 4 | TRUE |
| 2 | $z > 3$ |

g.c. TRUE

| $B$ | $C$ | *condition* |
|---|---|---|
| 4 | 1 | TRUE $\wedge$ $1 > 2$ |
| 2 | z | $z > 3 \wedge z > 2$ |

g.c. TRUE

Table V
THE RESULT OF $\pi_B(R_2)$ AND $\sigma_{C>2}(R_2)$

Table IV shows two example c-tables that we will use throughout this section. The left part of Table V shows the result of $\pi_B(R_2)$. The following theorem describes the complexity of the projection operator.

*Theorem 10:* The projection operator takes $O(s)$ time, where $s$ is the size of the c-table on which the projection is applied.

*Proof*: The operator goes though the c-tuples of the c-table exactly once and eliminates certain attributes. Therefore, the time complexity of the operator is equal to order the size of the c-table. ■

### B. Selection

*Definition 9 (syntax and semantics of selection):* We denote the selection over a c-table $T$ as $\sigma_\gamma(T)$, where $\gamma$ is a predicate formula over $\langle \mathbb{R}, \{>, =, +\} \rangle \cup \langle \mathbf{S}, \{=, \neq\} \rangle$ that references the variables $\{A_i\}_{i=1}^n$ that have the same names as the attributes of $T$. The pseudo-code for performing selection is presented in Algorithm 6.

**Algorithm 6** $\sigma_\gamma(T)$

```
for t ∈ T do
    θ(t) ← a substitution that substitutes every variable
           A_i with t[A_i] (the value for the attribute A_i in t).
    lc(t) ← lc(t) ∧ γ_θ(t)
end for
return T
```

*Theorem 11:* The selection operator is well defined.

*Proof*: We need to show that $rep(\sigma_\gamma(T)) \equiv \sigma_\gamma(rep(T))$ for every c-table $T$. But this is equivalent to proving that there exists valuations $v$ and $v'$ s.t. $v(\sigma_\gamma(T)) = \sigma_\gamma(v'(T))$. However, we have defined selection over c-tables in such a way so that $v(\sigma_\gamma(T)) = \sigma_\gamma(v(T))$ for any valuation $v$, which proves that selection is well defined. ■

The right part of Table V shows the result of $\sigma_{C>2}(R_2)$. The following theorem proves the time complexity of the selection operator.

*Theorem 12:* The selection, as we have defined it, takes $O(s * m)$ time, where $s$ is the size of the c-table and $m$ is the size of the selection condition.

*Proof*: The number of c-tuples in the c-table is bounded by $s$. For every c-tuple, we need to add to its local condition a condition of size $m$ and therefore the time complexity of the algorithm is $O(s * m)$. ■.

In order to save space, c-tuples with unsatisfiable local condition can be removed from the c-table, where we can use the *fast_simplify* algorithm to detect such c-tuples.

### C. Inner Join

*Definition 10 (syntax and semantics of inner join):* Consider a c-table $T_1$ with attributes $\{\bar{A}, \bar{B}\}$ and a c-table $T_2$ with attributes $\{\bar{B}, \bar{C}\}$. We denote the inner join of $T_1$ and $T_2$ on the set of attributes $\bar{B}$ as $T_1 \bowtie_{\bar{B}} T_2$. The pseudo-code for performing inner join is presented in Algorithm 7.

**Algorithm 7** $T_1 \bowtie_{\bar{B}} T_2$

```
1: T ← empty c-table with attributes Ā ∪ B̄ ∪ C̄
2: rename the variables in T2 so that T1 and T2 do no share
   variables
3: for t1 ∈ T1 do
4:    for t2 ∈ T2 do
5:       if fast_simplify(π_B̄(main(t1)) =
         π_B̄(main(t2)))! = false then
6:          main(t) ← (t1, π_C̄(t2))
7:          lc(t) ← lc(t1) ∧ lc(t2) ∧ (t1[B̄] = t2[B̄])
8:          add t to T
9:       end if
10:   end for
11: end for
12: gc(T) ← gc(T1) ∧ gc(T2)
13: return T
```

*Theorem 13:* The inner join operator is well defined.

*Proof*: Let $v$ be a valuation of the distinct variables of $T_1$ and $T_2$ (after Line 2 of Algorithm 7 is executed). We need to show that $v(T_1 \bowtie_{\bar{B}} T_2) = v(T_1) \bowtie_{\bar{B}} v(T_2)$. Let $t \in v(T_1 \bowtie_{\bar{B}} T_2)$. Then there must exist $t_1 \in T_1$ and $t_2 \in T_2$ such that the main part of $t$ is equal to the join of the main parts of $t_1$ and $t_2$. Then $t = v(t_1) \bowtie_{\bar{B}} v(t_2)$ and therefore $t \in v(T_1) \bowtie_{\bar{B}} v(T_2)$. The other direction is analogous. ■

| *A* | *B* | *C* | *condition* |
|---|---|---|---|
| 2 | $x$ | 1 | $x \neq 3 \wedge x = 4 \wedge$ TRUE |
| 2 | $x$ | z | $x \neq 3 \wedge x = 2 \wedge z > 3$ |
| 2 | 4 | 1 | TRUE $\wedge$ TRUE |

g.c. $x \neq 2 \wedge$ TRUE

Table VI
THE RESULT OF $R_1 \bowtie R_2$

Table VI shows the result of $R_1 \bowtie R_2$. The following theorem proves the time complexity of the inner join operator.

*Theorem 14:* Inner join takes $O(n' \cdot n'' \cdot (\sqrt{2})^m \cdot m^c)$ time, where $n'$ and $n''$ are the sizes of the c-tables that are being joined, $m$ is the size of the longest local condition in them, and $c$ is a constant.

*Proof*: The code inside the double `for`-loop is executed $n' \cdot n''$ number of times. The main time complexity in Lines 5-9 come from the call to the *fast_simplify* method, which takes $O((\sqrt{2})^m \cdot m^c)$ time to execute. ■.

### D. Union

*Definition 11 (syntax and semantics of union):* If $T_1$ and $T_2$ are c-tables, then we will denote their union as $T_1 \cup T_2$. The pseudo-code for calculating the union of c-tables is presented in Algorithm 8.

**Algorithm 8** $T_1 \cup T_2$

```
1: T ← empty c-table
2: rename the variables in T2 so that T1 and T2 do no share
   variables
3: for t1 ∈ T1 do
4:    add t1 to T
5: end for
6: for t2 ∈ T2 do
7:    add t2 to T
8: end for
9: gc(T) ← gc(T1) ∧ gc(T2)
10: return T
```

Note that we define the union operator to be duplicate preserving. Duplicate eliminating union can be performed by applying duplicate elimination to its result.

*Theorem 15:* The union operator is well defined.

*Proof*: We need to show that $v(T_1 \cup T_2) = v(T_1) \cup v(T_2)$ for any valuation $v$. Let $t \in v(T_1 \cup T_2)$. Then there exit c-tuple $t_1$ such that $t_1$ is in either $T_1$ or $T_2$ and $t = v(t_1)$. Therefore, $t \in v(T_1) \cup v(T_2)$. The reverse direction is analogous. ■.

The following theorem proves the time complexity of Algorithm 8.

*Theorem 16:* The time complexity of Algorithm 8 is $O(n + m)$ where $n$ and $m$ are the sizes of $T_1$ and $T_2$, respectively.

*Proof*: The time complexity of the algorithm comes from Lines 3-5, which take $O(n)$ time, and Lines 6-8, which take take $O(m)$ time. Therefore, the total time complexity of the algorithm is $O(n+m)$. ■.

### E. Monus

In bag relational algebra over bag relational tables monus is defined as: $T_1 \dot{-} T_2 = \{t_{[k]} | t \in T_1 \wedge k = max(count(t, T_1) - count(t, T_2), 0)\}$, where $t_{[k]}$ is used to denote the tuple $t$ replicated $k$ times and $count$ is a function that returns the number of occurrences of the tuple specified as the first parameter in the table specified as the second parameter. The following definition extends the monus operator to c-tables.

*Definition 12 (syntax and semantics of monus):* The monus of two c-tables $T_1$ and $T_2$ is defined as $T_1 \dot{-} T_2$. The pseudo-code for performing the monus operator is presented in Algorithm 9.

**Algorithm 9** $T_1 \dot{-} T_2$

1: rename the variables in $T_2$ so that $T_1$ and $T_2$ do no share variables
2: $V \leftarrow T_1$
3: $i \leftarrow 0$
4: **for** $t_1 \in T_1$ **do**
5: $\quad j \leftarrow 0$
6: $\quad$ **for** $t_2 \in T_2$ **do**
7: $\quad\quad X[i][j] = (main(t_1) = main(t_2)) \wedge lc(t_1) \wedge gc(T_1) \wedge lc(t_2) \wedge gc(T_2)$
8: $\quad\quad j \leftarrow j + 1$
9: $\quad$ **end for**
10: $\quad i \leftarrow i + 1$
11: **end for**
12: $gc(V) \leftarrow gc(V) \wedge \bigwedge_{j=1}^{m} [\bigvee_{i=1}^{n} (Y[1,j] = \cdots = Y[i-1,j] = Y[i+1,j] = \cdots = Y[n,j] = 0 \wedge Y[i,j] = 1)] \wedge \bigwedge_{i=1}^{n} [\bigvee_{j=1}^{m} (Y[i,1] = \cdots = Y[i,j-1] = Y[i,j+1] = \cdots = Y[i,m] = 0 \wedge Y[i,j] = 1)]$
13: **for** $t \in V$ **do**
14: $\quad lc(t) \leftarrow lc(t) \wedge \neg[\bigvee_{j=1}^{m} (X[i,j] \wedge (Y[i,j] = 1))]$
15: **end for**
16: **return** $V$

*Theorem 17:* The monus operator is well defined, where the definition of complete is changed to: $[Rep(T') \dot{-} Rep(T'')] \cup \{\emptyset\} \subseteq Rep(T' \dot{-} T'')$.

*Proof*: The algorithm first renames the variables of $T_2$ so that they are distinct from those in $T_1$. Next, it calculates the matrix $X$ and sets a restriction on the possible values for the matrix $Y$. The value of $X[i,j]$ contains the condition that must hold for the $i^{\text{th}}$ c-tuple of $T_1$ to be deleted from $T_1$ and the c-tuple that "deletes" it to be the $j^{\text{th}}$ c-tuple of $T_2$. The matrix $Y[i][j]$ has the restriction that for each $j$ there exists exactly one $i$ such that Y[i][j]=1 and that for each $i$ there exists exactly one $j$ such that $Y[i][j] = 1$ (the elements of the matrix $Y$ can only take the values 0 and 1). The matrix $Y$ is used to enforce the condition that every c-tuple $t_2$ in $T_2$ can be used to delete at most one c-tuple of $T_1$ and that every c-tuple $t_1$ in $T_2$ can be deleted at most once. Lastly, the local conditions that we add to the resulting c-table do the deletions. They specify that if for some valuation both $X[i][j]$ and $(Y[i][j] = 1)$ hold, (i.e., if a c-tuple $t'_i$ in $T'$ matches with a c-tuple $t''_j$ in $T''$ and the valuation is such that $t'_i$ can not be deleted by any c-tuple other then $t'_j$ and $t'_j$ can only delete $t'_i$), then the c-tuple that was constructed from the $i^{\text{th}}$ c-tuple in $T_1$ should be deleted from the resulting c-table $V$.

Given a valuation $v$, each c-tuple in $T_1$ will be deleted only if there exists a matching c-tuple in $T''$. Moreover, given a valuation $v$, every c-tuple in $T_2$ can delete at most one c-tuple from $T_1$. Therefore, the algorithm is correct and $Rep(T_1 \dot{-} T_2) \equiv [Rep(T_1) \dot{-} Rep(T_2)] \cup \{\emptyset\}$. Here $\{\emptyset\}$ is used to represent the empty c-table. ■

Note that we had to modify the definition of a complete relational algebra because we constructed the global condition of $T_1 \dot{-} T_2$ in such a way so that we allow for $\{\emptyset\}$ to be a possible representation. It is our believe that this is an intrinsic problem of monus when dealing with the closed world assumption.

A demonstration of how monus can be applied over the example c-tables from Table IV is shown in Table VII. The following theorem describes the time complexity of Algorithm 9.

| | |
|---|---|
| $x \neq 3 \wedge x \neq 2 \wedge$ $x = 4 \wedge$ TRUE $\wedge$ TRUE | $x \neq 3 \wedge x \neq 2 \wedge$ $x = 2 \wedge z > 3 \wedge$ TRUE |
| TRUE $\wedge\ x \neq 2 \wedge 4 = 4$ TRUE $\wedge$ TRUE | FALSE |

| *A* | *B* | *condition* |
|---|---|---|
| 2 | $x$ | $x \neq 3 \wedge \neg((X[1,1] \wedge$ $Y[1,1] = 1) \vee (X[1,2] \wedge Y[1,2] = 1))$ |
| 2 | 4 | TRUE $\wedge\ \neg((X[2,1] \wedge$ $Y[2,1] = 1) \vee (X[2,2] \wedge Y[2,2] = 1))$ |

g.c. $(x \neq 2) \wedge ((Y[1,1] = Y[2,2] = 1 \wedge Y[1,2] = Y[2,1] = 0) \vee (Y[1,2] = Y[2,1] = 1 \wedge Y[1,1] = Y[2,2] = 0))$

Table VII
SHOWS THE MATRIX $X$ AND THE RESULT FOR $R_1 \dot{-} R_2$

*Theorem 18:* Monus, takes $O(m \cdot n)$ time, where $m$ and $m$ are the sizes of the c-tables on which the operation is performed.

*Proof*: The complexity of the algorithm comes from the two `for`-loops. The code inside the double `for`-loops runs in constant time and it is executed $O(m \cdot n)$ time. ■

| $A$ | $B$ | *condition* |
|---|---|---|
| 2 | $x$ | $x \neq 3 \wedge (x \neq 4 \vee$ FALSE) |
| 2 | 4 | TRUE $\wedge (x \neq 4 \vee x = 3)$ |

g.c. $x \neq 2$

Table VIII
THE RESULT OF $\varepsilon(R_1)$

### F. Duplicate Elimination

The last relational algebra operation that we will explore in this section is duplicate elimination. In the relational case, duplicate elimination can be defined as a grouping on all the attributes. We adopt similar definition here.

*Definition 13:* We will denote the duplicate elimination operator applied to the c-table $T$ as $\varepsilon(T)$. We will compute $\varepsilon(T)$ using the formula $\varepsilon(T) = group_{\bar{A}}(T)$, where $\bar{A}$ are the attributes of $T$.

Note that the result of the group operation is a *nested c-table* (see Table IX for an example of a nested c-table). We define the semantics of a nested c-table and of the *group* operation in Section V-A.

*Theorem 19:* The duplicate elimination operator is well defined.

*Proof*: $\varepsilon(Rep(T)) \equiv group_{\bar{A}}(Rep(T)) \equiv Rep(group_{\bar{A}}(T)) \equiv Rep(\varepsilon(T))$, which proves that duplicate elimination is well defined. The fact that the equation $group_{\bar{A}}(Rep(T)) \equiv Rep(group_{\bar{A}}(T))$ holds follows from the fact that the $group$ operation is well defined over c-tables, which will be proven in Section V-A. ■

The result of $\varepsilon(R_1)$, where $R_1$ is the c-table defined in Table IV, is shown in Table VIII.

## V. APPLYING AGGREGATION TO C-TABLES

To the best of our knowledge, no research has been previously published in the area of applying grouping and aggregation to c-tables. We are aware of research on applying aggregation to fuzzy numbers ([23]) and to random variables ([24]), but the query results in these algorithms are approximations. On the other hand, the research done in constraint databases ([25]) has explored the problem of aggregation over constraint databases. Unfortunately, the operation of aggregation in most constraint database systems is not closed ([26]). We are also aware of recent research in the area of auditing confidential information ([27]), which, however, deals only with aggregation over Boolean variables.

In general, we would like to be able to evaluate a relational expression of the form ${}_{\bar{A}}\mathcal{F}_{agg_1(B_1),\ldots,agg_n(B_b)}T$, where $\bar{A} \cup \{B_i\}_{i=1}^{b}$ are the attributes of $T$, $\bar{A} = \{A_i\}_{i=1}^{a}$, and each $agg_i$ is one of the aggregates: *min*, *max*, *sum*, *count* and *avg*. In the relational case, the above expression is evaluated by grouping the tuples that have the same value for the attributes $\bar{A}$ into a single tuple that has this common value for the attributes $\bar{A}$. The value for the remaining attributes is calculated by applying the aggregation operations $\{agg_i\}_{i=1}^{n}$ to the value of the $\bar{B}$ attributes of the tuples in the group. In order to extend this definition to c-tables, we will need to be able to group c-tuples and perform aggregation on c-tuples.

### A. Grouping

The result of the grouping operation is a nested c-table that consists of nested c-tuples. An example nested c-table is shown in Table IX. Informally, a nested c-table consists of c-tuples that can have more than one value for some of the attributes. A formal definition follows.

*Definition 14 (nested c-tables and c-tuples):* A *nested c-tuple* with single valued attributes $\{A_i\}_{i=1}^{a}$ and multi-valued attributes $\{B_i\}_1^b$ is the sequence of mappings from $A_i$ to $D(A_i) \cup V_i$ for $i$ ranging from 1 to $a$ plus the sequence of mapping from $B_i$ to a bag of values over $D(B_i) \cup V_i$ for $i$ ranging from 1 to $b$ plus a local condition over $\langle \mathbb{R}, \{>, =, +\}\rangle \cup \langle \mathbf{S}, \{=, \neq\}\rangle$. Note that here $D(A)$ is used to denote the domain of $A$ and $V_i$ is used to represent a possibly infinite, but countable, set of variables over $D(A_i)$ in the first case and over $D(B_i)$ in the second case.

A *nested c-table* is a c-table that contains nested c-tuples. The semantics of a nested c-table is similiar to the semantics of a regular c-table as described in Section II-A (see Equations 1, 2, and 3). The only difference is that a nested c-table represents a set of nested bag relational tables (see [28]) under different valuations and consists of a bag of nested c-tuples.

The algorithm for performing the grouping uses the concept of a semi-unifiable c-tuples and the $\prec$ relation for c-tuples, which are formally presented next.

*Definition 15 (semi-unifiable c-tuples):* The c-tuples $\{t_i\}_{i=1}^{n}$ of the c-table $T$ are semi-unifiable relative to the set of attributes $\bar{A}$ exactly when the expression $\bigwedge_{i,j=1}^{n} \pi_{\bar{A}}(main(t_i)) = \pi_{\bar{A}}(main(t_j))$ is satisfiable under some valuation.

Informally, a bag of c-tuples are semi-unifiable relative to the set of attributes $\bar{A}$ exactly when the c-tuples can be potentially grouped into a single nested c-tuple in which $\bar{A}$ are the single-valued attributes.

*Definition 16 (the $\prec$ relation):* We will write $t_1 \prec_A t_2$, where $t_1$ and $t_2$ are c-tuples and $\bar{A}$ is a set of attributes exactly when $main(\pi_{\bar{A}}(t_2))$ can be constructed from $main(\pi_{\bar{A}}(t_1))$ by substituting some of the variables in $main(\pi_{\bar{A}}(t_1))$ with constants.

Informally, the $\prec$ relation compares the main parts of two c-tuples to determine if one c-tuple has more specific values than the other. The $\prec$ relation is transitive and defines partial order.

*Definition 17 (syntax and semantics of grouping):* We denote the result of grouping by the attributes $\bar{A}$ of $T$ as

| $A$ | $B$ | *condition* |
|---|---|---|
| 2 | $x$ | $x \neq 3 \wedge (x \neq 4 \vee \text{FALSE})$ |
| 2 | 4 | TRUE $\wedge$ $(x \neq 4 \vee x = 3)$ |
| 2<br>2 | 4 | $x \neq 3 \wedge$ TRUE $\wedge$ $x = 4$ |

g.c. $x \neq 2$

Table IX
THE RESULT OF $group_B R_1$

| $A$ | $B$ | $C$ | condition |
|---|---|---|---|
| $x$ | $y$ | 1 | $(x + y = 3) \vee (x > 4) \vee (x < 0)$ |
| $x$ | 3 | 2 | $(x + y = 3 \wedge x < 2) \vee (x < 0)$ |
| 2 | 3 | 3 | $x > 5$ |
| 2 | 3 | 4 | TRUE |
| 3 | 4 | 5 | TRUE |

Table X
EXAMPLE C-TABLE $R$

$group_{\bar{A}}(T)$. The pseudo-code for performing the grouping operator is shown in Algorithm 10.

The result of the grouping operation will be a *nested c-table*, that is, the value of a field in it may be a bag of values. For example, in $group_{\bar{A}}(T)$ the values for the attributes in $\bar{A}$ will be single values and for the the rest of the attributes - bag of values. The result of $group_B R_1$ is shown in Table IX, where $R_1$ is shown in Table IV.

*Theorem 20:* The *group* operator is well defined, that is, $group_{\bar{A}}(Rep(T)) \equiv Rep(group_{\bar{A}}(T))$.

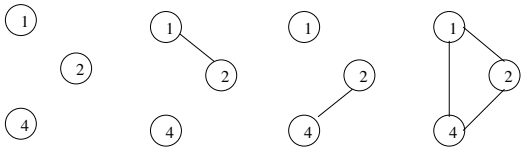
*Proof*: Line 1 copies $T$ into the resulting c-table (our running example is on the the c-table $R$ shown in Table XVIII and we show how to calculate $group_{A,B}(R)$).

Line 2 clusters the c-tuples into e-bags relative to the attributes of $\bar{A}$. Table XI shows the two e-bags that will

| $A$ | $B$ | $C$ | *condition* |
|---|---|---|---|
| $x$ | $y$ | 1 | $((x + y = 3) \vee (x > 4) \vee (x < 0)) \wedge t = 1$ |
| $x$ | 3 | 2 | $(x + y = 3 \wedge x < 2) \vee (x < 0)$ |
| 2 | 3 | 3 | $x > 5$ |
| 2 | 3 | 4 | TRUE |

| $A$ | $B$ | $C$ | *condition* |
|---|---|---|---|
| $x$ | $y$ | 1 | $((x + y = 3) \vee (x > 4) \vee (x < 0)) \wedge t \neq 1$ |
| 3 | 4 | 5 | TRUE |

Table XI
E-BAGS IN $group_{A,B}(R)$




Graph 1 Graph 2 Graph 3 Graph 4

Figure 2. The four possible graphs for the first e-bag

| $C[i]$ | value | c-tuples |
|---|---|---|
| $C[1]$ | $x < 0 \wedge t = 1$ | $\{1, 2, 4\}$ |
| $C[2]$ | $0 \leq x < 2 \wedge x + y = 3 \wedge t = 1$ | $\{1, 2, 4\}$ |
| $C[3]$ | $2 \leq x \leq 4 \wedge x + y = 3 \wedge t = 1$ | $\{1, 4\}$ |
| $C[4]$ | $4 < x \leq 5 \wedge t = 1$ | $\{1, 4\}$ |
| $C[5]$ | $x > 5 \wedge t = 1$ | $\{1, 3, 4\}$ |
| $C[6]$ | $x < 0 \wedge t \neq 1$ | $\{2, 4\}$ |
| $C[7]$ | $0 \leq x < 2 \wedge x + y = 3 \wedge t \neq 1$ | $\{2, 4\}$ |
| $C[8]$ | $x > 5 \wedge t \neq 1$ | $\{3, 4\}$ |
| $C[9]$ | $(x + y \neq 3 \wedge 0 \leq x \leq 4) \vee (4 < x \leq 5 \wedge t \neq 1)$ | $\{4\}$ |

| $C[i]$ | value | c-tuples |
|---|---|---|
| $C[1]$ | $((x + y = 3) \vee (x > 4) \vee (x < 0)) \wedge t \neq 1$ | $\{1, 2\}$ |
| $C[2]$ | $((x + y \neq 3) \wedge (0 \leq x \leq x)) \vee (t = 1)$ | $\{2\}$ |

Table XII
THE ARRAY $C$ FOR THE TWO E-BAGS

| $D[i]$ | values | c-tuples |
|---|---|---|
| $D[1]$ | $C[1] \vee C[2]$ | $\{1, 2, 4\}$ |
| $D[2]$ | $C[3] \vee C[4]$ | $\{1, 4\}$ |
| $D[3]$ | $C[5]$ | $\{1, 3, 4\}$ |
| $D[4]$ | $C[6] \vee C[7]$ | $\{2, 4\}$ |
| $D[5]$ | $C[8]$ | $\{3, 4\}$ |
| $D[6]$ | $C[9]$ | $\{4\}$ |

| $D[i]$ | values | c-tuples |
|---|---|---|
| $D[1]$ | $C[1]$ | $\{1, 2\}$ |
| $D[2]$ | $C[2]$ | $\{2\}$ |

Table XIII
THE ARRAY $D$ FOR THE TWO E-BAGS

be constructed after applying Line 2 to our example. Note that Line 2 is equivalence preserving and that c-tuples from different e-bags cannot contribute to the same resulting nested c-tuple under any valuation. This is why it suffices to perform the *group* operation to the c-tuples in each e-bag and then merge the results.

Line 3 partitions each e-bag further into r-bags. In other words, we partition the space over which the local conditions of the c-tuples in the e-bags is defined into non-overlapping polyhedra. Each r-bag corresponds to a set of disjoint

| $A$ | $B$ | $C$ | *condition* |
|---|---|---|---|
| $x$ | $y$ | 1 | $y \neq 3 \wedge x \neq 2 \wedge R$ |
| $x$ | 3 | 2 | $y \neq 3 \wedge x \neq 2 \wedge R$ |
| 2 | 3 | 4 | $y \neq 3 \wedge x \neq 2 \wedge R$ |
| $x$ | $y$ | 1<br>2 | $x \neq 2 \wedge y = 3 \wedge R$ |
| 2 | 3 | 4 | $x \neq 2 \wedge y = 3 \wedge R$ |
| $x$ | $y$ | 1 | $x = 2 \wedge y \neq 3 \wedge R$ |
| 2 | 3 | 2<br>4 | $x = 2 \wedge y \neq 3 \wedge R$ |
| $x$ | $y$ | 1<br>2<br>4 | $x = 2 \wedge y = 3 \wedge R$ |

$R = ((x < 0 \wedge t = 1) \vee (0 \leq x < 2 \wedge x + y = 2 \wedge t = 1))$

Table XIV
THE CONTRIBUTION OF THE FIRST R-BAG OF THE FIRST E-BAG TO THE RESULT OF $group_{A,B}(R)$

**Algorithm 10** $group_{\bar{A}}(T)$

1: $V \leftarrow T$

2: Cluster the c-tuples of $V$ into biggest bags of semi-unifiable c-tuples relative to $\bar{A}$ - we will call this *e–bags*. If a c-tuple belongs to more than one e-bag, then make copies of the c-tuple and put a copy in each e-bag. To do so, add the local condition $x = i$ to the $i^{\text{th}}$ copy of the c-tuple for $i < u$ and the local condition $\bigwedge_{i=1}^{u-1} x \neq i$ to the $u^{\text{th}}$ copy, where $x$ is a newly introduced variable and $u$ is the number of times the c-tuple is copied.

3: Partition each e-bag further into r-bags. To do so, call *break_up*$(\bigvee_{i=1}^{p} lc(t_i))$, where $\{t_i\}_{i=1}^{p}$ are the c-tuples in the e-bag that is being processed. This will produce a set of non-overlapping conjunctions $\{c_i\}_{i=1}^{w}$. Let $C = \{c_i\}_{i=1}^{w}$. Rewrite the local condition of each $t_i$ as a disjunction of $c_i$s. Next, break $C$ into equivalence classes relative to the operation $\sim$. We define $c_i \sim c_j$ exactly when the set of the rewritten local conditions in which the two conjunctions appear is the same. Next, create an array $D$, where $D[i]$ is the disjunction of all the conjunctions in the $i^{\text{th}}$ equivalence class. Reconstruct $V$ by substituting each e-bag with a bag of *r-bags*. The c-tuples in $i^{\text{th}}$ r-bag of a given e-bag will have the same local condition as the corresponding value of $D[i]$ and the main parts will correspond to the c-tuples that contained the local conditions that formed the equivalence class corresponding to $D[i]$.

4: From each r-bag, create a set of vertices, where each vertex corresponds to a distinct c-tuple in the r-bag (i.e., for duplicate c-tuples we will have a single vertex). Next, find all spanning undirected graphs that are transitive and have the property that if there is an edge between the vertices $n_1$ and $n_3$ and there exists a third vertex $n_2$ such that $t_1 \prec_{\bar{A}} t_2$ and $t_2 \prec_{\bar{A}} t_3$, where $t_1$, $t_2$ and $t_3$ are the c-tuples corresponding to the vertices, then there are edges between $n_1$ and $n_2$ and between $n_2$ and $n_3$.
Next, the set of nested c-tuples that correspond to each graph are created. Their union yields the result of doing the grouping. More precisely, suppose that we are examining an r-bag $r$ and a graph $G$ associated with it. Since $G$ is transitive, it will contain a set of disjoint complete sub-graphs, where each such sub-graph will correspond to a resulting nested c-tuples. If the vertices in the complete sub-graph belong to the c-tuples $\{t_i\}_{i=1}^{p}$, then the corresponding nested c-tuple will have the single value $(x_1, \ldots, x_a)$ for the attributes $\bar{A}$, the bag of values $\{| \pi_{\bar{B}} main(t_i) |\}_{i=1}^{p}$ for the attributes $\bar{B}$, and the local condition $L_r \wedge R_G \wedge (\bigwedge_{i=1}^{p} [(\pi_{\bar{A}} main(t_i)) = (x_1, \ldots, x_a)])$. The condition $L_r$ is the local condition of the r-bag $r$. The condition $R_G$ is the condition that projection on the $\bar{A}$ attributes of the main parts of the c-tuples that correspond to nodes in $G$ that are connected should be equal, while the projection on the $\bar{A}$ attributes of the main parts of the c-tuples that correspond to nodes in $G$ that are not connected should be distinct.

polyhedra. Note that this operation is equivalence preserving. The additional constraint that all the conjunctions that form the $D[i]$ of a given r-bag appear in the same set of c-tuples' local conditions guarantees that the r-bags partition the possible valuations, that is, under every valuation the local condition of at most one r-bag of every e-bag will be true. In other words, given an arbitrary valuation and an e-bag of r-bags, either none of the c-tuples' local conditions will be true or the local conditions of all the c-tuples in exactly one r-bag will be true. For our example, Tables XII and XIII show the value of the $C$ and $D$ array, respectively. Note that, in order to keep the example simple, the local conditions are not normalized using the algorithm from [18].

Next, the algorithm constructs a set of graphs for each r-bag, where each graph corresponds to a valuation. In a graph, there is an edge between two vertices if under the corresponding valuation it is true that $\pi_{\bar{A}}(main(t')) = \pi_{\bar{A}}(main(t''))$, where $t'$ and $t''$ are the c-tuples corresponding to the vertices. A graph is valid, that is, a corresponding valuation exists exactly when (1) the graph is transitive (2) if there is an edge between the vertices $n_1$ and $n_3$ and there exists a third vertex $n_2$ such that $n_1 \prec_{\bar{A}} n_2$ and $n_2 \prec_{\bar{A}} n_3$, then there are edges between $n_1$ and $n_2$ and between $n_2$ and $n_3$. This is why all the graphs having these two properties are constructed and these graphs show which c-tuples in the r-bag will be grouped relative to the attributes $\bar{A}$ under different valuations. Figure 2 shows the graph for the first r-bag of the first e-bag, where the c-tuple numbers are preserved from Table XI. The resulting c-tuples that are constructed from the four possible graphs are shown in Table XIV. ■

The following theorem proves the time complexity of Algorithm 10.

*Theorem 21:* A *variable c-tuple* is a c-tuple that has variables in it it, while a *regular c-tuple* is a c-tuple that does not. Let $v$ be the number of variable c-tuples in $T$, $m$ be the greater of the size of the longest c-tuple and the size of the global condition of $T$, $n$ be the number of regular c-tuples with distinct main parts, $r$ be the highest count of regular c-tuples that have the same main part but distinct local conditions, $s$ be the number of attributes in $T$, and $c$ be a constant. Then the total time to perform Algorithm 10

is $O((2^v+n)\cdot s+(2^v+n)\cdot 3^{\sqrt{2}^{m\cdot(v+r)}}\cdot(m\cdot(v+r))^c+(2^v+n)\cdot 2^{v+r}\cdot 2^{v+n})$.

*Proof*: Line 2 of Algorithm 10 takes $O(2^v+n)\cdot s$ time because it may take as much as $O(2^v\cdot s)$ time to partition the variable c-tuples and then $O(n\cdot s)$ time to determine the groups for the regular c-tuples. Note that we get this low time bound thanks to the fact that regular c-tuples with distinct main parts can not appear in the same e-bag.

Line 3 will take $O((2^v+n)\cdot 3^{\sqrt{2}^{m\cdot(v+r)}}\cdot(m\cdot(v+r))^c)$ time because the size of a c-tuple's local condition may grow to a size of $O(m\cdot(v+r))$ after the normalization procedure from [18] is applied.

Line 4 takes $O((2^v+n)\cdot 2^{v+r}\cdot 2^{v+n})$ time because there maybe as much as $2^{v+r}$ r-bags in each e-bag and each r-bag may contain as much as $n+v$ distinct c-tuples and therefore there are $2^{v+n}$ possible graphs for each r-bag. ■

### B. Performing the Aggregation

Now that we have defined how grouping over c-tables can be done, performing aggregation is straightforward. The following definition contains the details.

*Definition 18 (syntax and semantics of aggregation):* Let $T$ be a c-table. We will denote a grouping by the attributes $\bar{A}=\{A_i\}_{i=1}^a$ and aggregation for the attributes $\bar{B}=\{B_i\}_{i=1}^b$ as ${}_{A_1,\ldots,A_a}\mathcal{F}_{agg_1(B_1),\ldots,agg_b(B_b)}T$, where the sets $\bar{A}=A_1,\ldots,A_a$ and $\bar{B}=B_1,...,B_b$ are disjoint and their union yields all the attributes in $T$. The value for *agg* can be *min*, *max*, *sum*, *count*, or *avg*. Algorithm 11 shows the pseudo-code for performing the aggregation.

**Algorithm 11** ${}_{A_1,\ldots,A_a}\mathcal{F}_{agg_1(B_1),\ldots,agg_b(B_b)}T$

1: $V\leftarrow group_{\bar{A}}T$
2: **for** $t$ in $V$ **do**
3: perform the mapping from Table XV to Table XVI on $t$, where $\{x_i\}_i=1^n$ are new variables and the function *con* is defined in Table XVII.
4: **end for**
5: **return** $V$

Note that Line 3 of Algorithm 11 performs aggregating over the $\bar{B}$ attributes by introducing new variables in the main parts of the result and moving the aggregations to the local conditions.

*Theorem 22:* The aggregation operator that is defined in Algorithm 11 is well defined.

*Proof*: The correctness of the grouping algorithm follows from Theorem 20 and the correctness of the *con* operator. Let us next examine the *con* operator. For the *min* operation it adds the condition that the new variable must be smaller than the value for the other c-tuples for that attribute in the group, which is the desirable behavior. The correctness of the *max* operation is analogous. The *count* operation is implemented correctly because it returns the count of the c-tuples in each group. The *sum* operation adds the condition that the value for the aggregate attribute must be equal to the sum of the values for that attribute in each group, which is the expected behavior. Finally, for the *avg* operator we add the condition that $x*n$ must be equal to the sum of the values for the aggregate attribute and therefore the new value for the attribute will be $\frac{\sum_{i=1}^{n} b_i}{n}$, which is exactly the average operator. ■

Table XVIII shows the result of ${}_B\mathcal{F}_{sum(A)}R_1$, where Table $R_1$ is defined in Table IV

| $A$ | $B$ | *condition* |
|---|---|---|
| $a_1\ldots a_k$ | $b_1^1\ldots b_n^1$ | $c$ |
| | $\ldots$ | |
| | $b_1^p\ldots b_n^p$ | |

Table XV
A COMPLEX C-TUPLE $t$

| $A$ | $B$ | *condition* |
|---|---|---|
| $a_1\ldots a_k$ | $x_1\ldots x_n$ | $c\wedge con(x_1,agg_1,b_1^1,\ldots,b_1^p)\wedge$ $\cdots\wedge con(x_n,agg_n,b_n^1,\ldots,b_n^p)$ |

Table XVI
THE RESULT OF ${}_{\bar{A}}\mathcal{F}_{agg_1(B_1),\ldots,agg_n(B_n)}(t)$

## VI. CONCLUSION AND FUTURE RESEARCH

In the paper, we presented algorithms for querying c-tables extended with linear conditions using the closed world assumption. We have chosen this representation because it is the least expressive extension of c-tables over which bag relational algebra with grouping and aggregation is closed

| $(agg)$ | $con(x,agg,b_1,\ldots,b_n)$ |
|---|---|
| $min$ | $\bigwedge_{i=1}^{n}(x\le b_i)$ |
| $max$ | $\bigwedge_{i=1}^{n}(x\ge b_i)$ |
| $count$ | $n$ |
| $sum$ | $x=\sum_{i=1}^{n} b_i$ |
| $avg$ | $\underbrace{x+\cdots+x}_{n\text{ times}}=\sum_{i=1}^{n} b_i$ |

Table XVII
EXPLAINS THE OPERATOR *con*

| $A$ | $B$ | *condition* |
|---|---|---|
| 2 | $x$ | $x\neq 3\wedge(x\neq 4\vee$ FALSE$)$ |
| 2 | 4 | TRUE $\wedge\ (x\neq 4\vee x=3)$ |
| y | 4 | $x\neq 3\wedge$ TRUE $\wedge\ x=4\wedge y=2+2$ |

g.c. $x\neq 2$

Table XVIII
THE RESULT OF ${}_B\mathcal{F}_{sum(A)}R_1$

and can be well defined. As expected, the running time of the presented algorithms is polynomial relative to the size of the certain information and non-polynomial relative to the size of the incomplete information.

A major topic for future research is optimizing the algorithms for performing the different relational operations. For example, in the relational case, the join between two tables can be performed in different ways and the efficiency of performing the join depends on the implementation. The same applies for joining c-tables.

In general, there are different ways of performing the deferent relational algebra operations over c-tables. The purpose of this paper is to define their semantics by presenting example algorithms for doing the operations. The presented algorithms are not optimal and optimization techniques such as dynamic programming and iterative dynamic programming can be used to optimize the different relational algebra operators over c-tables.

Other possible extensions of the presented work follow.

• Explore c-tables with variables over additional domains, such as date or currency. • Speed up the presented algorithms by sacrificing their accuracy, that is, explore approximate query answering for incomplete information.
• Explore integrity constraints for incomplete information and how they can be used to perform semantic query optimization.
• Extend research done in relational databases, such as research on view maintenance, transaction control, logging and recovery, to databases that contain incomplete information.
• Explore introducing an ordering of the c-tuples in a c-table and defining an *order by* operator.

## REFERENCES

[1] L. Stanchev, "Querying Incomplete Information using Bag Relational Algebra," *eKNOW 2010, Second International Conference on Information Process, and Knowledge Management*, pp. 110–119, 2010.

[2] T. Imielinski and W. Lipski, "Incomplete Information in Relational Databases," *Journal of Association of Computing*, vol. 31, no. 4, pp. 761–791, October 1984.

[3] L. Libkin and L. Wong, "Some Properties of Query Languages for Bags," *Proceedings of Database Programming Languages*, pp. 97–114, 1994.

[4] G. Grahne, *The problem of Incomplete Information in Relational Databases*. Berlin: Springer-Verlag, 1991.

[5] R. Reiter, "A Sound and Sometimes Complete Query Evaluation Algorithm for Relational Databases with Null Values," *JACM*, vol. 33, no. 2, pp. 349–370, 1986.

[6] L. Y. Yuan and D.-A. Chiang, "A sound and Complete Query Evaluation Algorithm for Relational Databases with Null Values," *ACM*, 1988.

[7] L. Libkin, "Query Language Primitives for Programming with Incomplete Databases," *Proceedings of DBPL*, 1995.

[8] P. Buneman, A. Jung, and A. Ohori, "Using Powerdomains to Generalize Relational Databases," *Theoretical Computer Science*, vol. 91, no. 1, 1991.

[9] R. Reiter, *On closed world databases, Logic and databases*. Plenum Press, 1978.

[10] G. Grahne, "Dependency Satisfaction in databases with Incomplete Information," *Proceedings of International Conference on Very Large Data Bases*, pp. 37–45, 1984.

[11] J. A. Biskup, "A Formal Approach to null Values in Database Relations," *Advances in Database Theory*, pp. 299–341, 1981.

[12] E. F. Codd, "Understanding Relations (Installment 7)," *FDT Bull. of ACM-SIGMOD*, vol. 3, no. 4, pp. 23–28, December 1975.

[13] ——, "Extending the Database Relational Model to Capture more Meaning," *ACM Transactions on Database Systems*, vol. 4, no. 4, pp. 397–434, December 1979.

[14] J. Grant, "Null values in Relational Data Base," *Information Processing Letters*, vol. 6, no. 5, pp. 156–157, October 1977.

[15] T. Imielinski and W. Lipski, "On Representing Incomplete Information in a Relational Data Base," *Proceedings of the 7th International Conference on Very Large Data Bases*, pp. 388–397, September 1981.

[16] M. Fischer and M. O. Rabin, "Super Exponential Complexity of Presburger Arithmetic," *Project MAC Tech. Mem. 43. MIT*, 1974.

[17] S.Basu, "New Results on Qunatifier Elimination over Real Closed Fields and Applications to Constraint Databases," *JACM*, vol. 46, no. 4, pp. 537–555, 1999.

[18] J. L. Lassez and K. McAloon, "Applications of a Canonical Form of Generalized Linear Constraints," *Journal of Symbolic Computation*, vol. 13, pp. 1–24, 1992.

[19] J. L. Lassez, T. Huynh, and K. McAloon, "Simplification and Elimination of Redundant Arithmetic Constraints," *Proceedings of NACLP*, 1989.

[20] A. Tarski, "A Decision Method for Elementary Algebra and Geometry," *University of California Press*, 1951.

[21] F. Jerrante and C. Rackoff, " A Decision Procedure for the First Order Theory of Real Addition with Order," *SIAM Journal of Computing*, vol. 4, no. 1, pp. 69–76, 1975.

[22] D. Kossmann and K. Stocker, "Iterative Dynamic Programming: A New Class of Query Optimization Algorithms," *ACM Transactions on Database Systems*, vol. 25, no. 1, 2000.

[23] G. Klir, U. Clar, and B. Yuan, *Fuzzy Set Theory. Foundations and Applications*. Prentice Hall, 1997.

[24] M. D. Springer, "The Algebra of Random Variables," *Wiley series in probability and mathematical statistics*, 1979.

[25] G. Kuper, L. Libkin, and J. Paredaens, *Constraint Databases*. Springer, 1998.

[26] G. M. Kuper, "Aggregation in Constraint Databases," *Proceedings of the 1st International Workshop on Principles and Practice of Constraint Programming*, pp. 161–172, 1993.

[27] J. Kleinberg, C. Papadimitriou, and P. Raghavan, "Auditing Boolean Attributes," *PODS*, pp. 86–91, 2000.

[28] A. Makinouchi, "A Consideration of Normal Form of Non-necessarily-normalized Relations in the Relational Data Model," *Proceedings. of International Conference on Very Large Data Bases*, pp. 447–453, 1977.

# Complex Navigation Systems - Some Issues and Solutions

Vladislav Martínek and Michal Žemlička
*Dept. of Software Engineering*
*Charles University in Prague*
*Prague, Czech Republic*
*martinek@ksi.mff.cuni.cz, zemlicka@ksi.mff.cuni.cz*

***Abstract*—Navigation is a kind of application widespread especially in mobile devices. We can find complex navigation systems that combine two or more types of navigation. This can lead to a complex service which can increase the efficiency and comfort of user's movement. The user and even the path can be characterized by a large number of variables. This fact opens a problem of finding a path that will satisfy chosen variables. We tried to point out interesting issues concerning combined navigation. For some of the issues, we propose possible solutions. The solutions are based either on solutions from existing products or on our own experience from the JRGPS project supporting combination of public transport and walk. The path reliability is one of the important factors for connection planning. We propose several different points of view at path reliability under the condition of public transport network. Using detailed data from the connection provider, we can see how the characteristic features of public transport networks involve this path parameter.**

***Keywords*-navigation; complex navigation; pedestrian navigation; public transportation; mobile device; GPS device**

## I. Introduction

Path searching applications (navigation systems) can be built in cars, they can be run on mobile devices to navigate walkers or even bikers. They can be used for searching connection in timetables of public transport networks. The usability in human transportation is various.

There are two main kinds of navigation. First, there are applications, where the movement is depending mostly on the user itself (walkers, cars, bikers and so on). Second, there are applications dealing with the scheduled movement while using different transport services. The border between scheduled movement and individual movement do not have to be always strict. We suppose that good navigation system should be able to combine both the above mentioned approaches. Furthermore, there can be restrictions on the path found (reliability, safety, price, usability for people with limited movement abilities, or given movement speed). The good navigation should also take into account individual preferences and limitations of the user in the context of the planned path.

Complex navigation systems combine several types of navigation to increase the overall effectiveness and comfort of users movement. Providing complex path planning leads to the problems of connection between transport networks.

This paper is based on our experience in the development of an application for mobile devices that search the shortest path combining public transport and walking. We worked with timetables and map base for the city of Prague. This paper is an extended version of a paper [1].

In the following text, we will outline several problems and issues connected with developing complex navigation. We will also propose solutions used in our prototype application. All the screenshots introduced in the paper are from this application. Finally, we will summarize advantages and disadvantages of combined navigation and propose intention of our future work.

## II. State of Art

In the area of pedestrian navigation used in urban location, there are at least two different approaches. The first one is a tourist guide which is able to plan point-to-point path using walk exclusively. While navigating through the path, the navigation is able to highlight points of interest along the path [2]. The second one is a navigation that combines the walk with other types of transportation suitable for pedestrians. This application can still serve as a tourist guide, but the aim is often to serve as a path planning tool for everyday use.

On the other hand, the navigation systems can be categorized according to their dependency on a connection to a mobile operator. The off-line navigation system is able to work in areas where the connection is not available. For example, it can happen in a subway train going between two neighbor stations. The path plan is computed by the mobile device itself in off-line navigation. The on-line navigations system can have more actual data than the off-line ones. The cost of data transfers from the operator can increase the cost of on-line navigation.

The complex navigation systems, that combine several types of transportation of pedestrians, are often implemented as on-line services. For example, "Google Maps Navigation" [3] or "Navitime" [4] are a kind of on-line services. On the other hand, "Nokia Maps" [5] provides an off-line navigation for pedestrians, but the complexity is so far limited for a combination of car and walk. The situation is similar in "TomTom" [6] , "Navigon" [7] , "Mio" [8].

## III. ISSUES

When creating complex navigation, there are several problems that should be challenged to reach certain quality of resulting application.

### A. Navigation in a Scheduled Network

The services in a public transport (scheduled) network should follow a valid timetable. The separate parts of path in this network are therefore predetermined in space and time. These parts can have a fixed starting moment or they can start periodically.

Path planning in a scheduled network has one important property: The plan of the entire path can vary in space according to the starting time value. The consequence is that the path plan may differ significantly for two relatively close moments. See Figure 1.

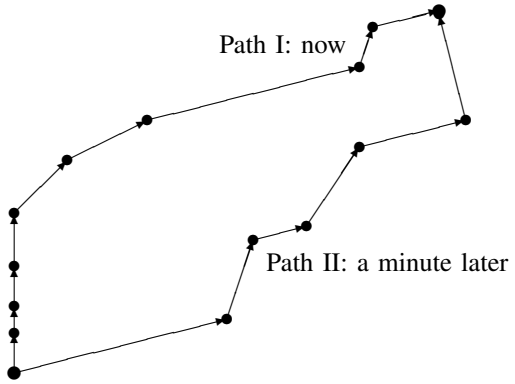



Figure 1. Path Plan Variability: The requirements for the highlighted path plans differ only by the starting time.

The validity of timetables is limited and should be kept up. In addition, there may be unplanned changes in the schedules. Temporary exceptions, technical problems, and other unforeseen events may affect the schedule as well.

### B. Navigation of Walkers

It is not necessary to consider the current time when planning the walking route as a walker can typically start the journey at any time. But some sections of the walking path may be passable only at certain times of the day or in some days only. In this case, the current time should be considered. It is possible to find more complicated cases where the passage of some sections may be significantly more difficult and slower in certain periodical moments.

### C. Combination of Different Navigations

*1) Fixed and Free Sections of Path:* When planning the combined path, it is necessary to distinguish sections fixed in time and free sections. The time gaps can appear between the parts of a planned path. These gaps can represent, for example, waiting for the service of public transport. The free parts of planned path could be moved in order to minimize the gaps or to satisfy the preferences of the user.

In order to properly combine the sections of path, it is necessary to know their length. In addition, the starting point for fixed sections is given and cannot be moved. The time is the key parameter for planning of the combined path. Individual public transport services are represented by fixed sections only. Walking paths could be represented by both fixed and free sections. It is necessary to know the time needed to get through the section.

When planning the combined path, it is necessary to determine the length of each section before it is planned into the path. In the case of public transport, the duration is determined by the current time and the valid timetable of a service which implements current section. The duration of walk section is determined according to the time needed by the user to get through the section. In both cases, the user defined parameters will be important for the planning. These parameters will affect the choice of sections, and in the case of walk, they will also affect the length of individual sections.

Additional information about the character of a section is needed in the case of more complicated sections of walk path such as barriers or super elevated parts of path. The time to overcome such a section is based on current dispositions of a particular user.

The main parameter for planning paths in public transport is time. Walking paths are often based on distance parameter in tourist navigators. The distance can be easily converted to time on simple walking sections. In more complex sections such as barriers or various sections of elevation, we require additional information about the section. Time needed to pass such a section is a much more practical information for the planning. Walking around a city is completely different from the normal tourist routes and the duration of the travel is incomparably more important than the distance.

*2) Combination of Different Search Networks:* Both public transport network and a network of pedestrian pathways may be very large and when combined, the total size of the searched network can grow over computation possibilities of mobile devices. Effectiveness of the overall planning is strongly dependent on the chosen solution.

Some solutions of searching the shortest path are compared in [9]. For the combination of different networks, the approach similar to “highway hierarchy” [10] seems to be promising.

Currently, portable devices have sufficient computing and memory capacity to handle the combination of path planning. It is necessary to choose an appropriate representation of data that will not exhaust the memory capacity of portable devices. This can lead to an application independent on the current availability of connection. Update of the timetables can be made when the connection or other mechanism for the update is available.

### D. Path Reliability

Beside the path length, the reliability of path found might be the important parameter which involves usability of the path. To determine the reliability of public transport services, we may require an additional data from the carrier. On the other hand, the reliability of the connection can be viewed as the frequency of services at the particular section.

Some fixed sections can be repeated in a relatively short periods of time. This behavior is similar to the behavior of free sections. For example, tube in the rush hour when the arrivals are relatively frequent. If the user misses such planned service, it does not have to be necessary to re-plan the entire path.

An interesting challenge might be planning a path, where the user may miss some of the services or even all of them. Respectively, missing a service will lead to a minimal delay in following path sections.

When dealing with unreliable network, one of the approaches is to use approximative methods of finding optimal path. Two such methods are compared in [11].

*Response to Failures in the Network:* The failures of certain parts of network can occur in both public transport network and network of walk paths. Downtime can be known in advance and then it should be included in the update.

The failure may occur suddenly and then the user should have a possibility to change planning options to bypass actually unreachable part of network.

### E. Appropriate Map Data

One of the major problems is the unavailability of appropriate map data for the planning of pedestrian paths. Most of the existing map data is not sufficiently detailed to pedestrians. First, the map base should include actual data for pedestrians – sidewalks, crosswalks, pedestrian zones, footpaths and other routes applicable to pedestrians. Spatial data should be in vector format, which can be easily used to create search structures.

*1) Crosswalks:* The map data should contain details of the crosswalks, or crosswalks with the signaling device where it is necessary to calculate the specific interval for the path section. The crosswalks are essential for legal crossing of the roads. In most countries, pedestrians are not allowed to cross the road in places close to the crosswalks. The situation is more complicated when the sidewalk is bordered by a railing or other such barrier. To determine where the crossing of the road is acceptable outside the crosswalk is a separate problem. However, if the map data are not detailed enough, we cannot solve this problem anyway.

*2) Spatial Data:* The map data should contain information about elevation. Due to the variable dispositions of walkers the map data should contain information about various barriers. It is particularly important to distinguish the high thresholds of sidewalks or stairs, because for some user, it is an obstacle, for some users it is not.

*3) Grade-separated Crossings:* Important information is the grade-separated crossings like bridges and subways. In the cases, where it is not possible to move freely between the levels of path, the incorrectly labeled crossings could lead to mistakes in navigation. The various levels of path need to be distinguished also when entering the starting point.

This information is relevant for the navigation in public transport itself. The path between refuges of one stop is often realized by a special crossing, like underpass or stairway. It is necessary to know the properties of the path connecting these refuges to determine correctly the length of the transfer within the stop.

### F. Combining Pedestrian and Public Transport Networks

Another problem is the combination of data for pedestrians and public transport network. Combination must be done in both directions.

*1) Street Refuges in the Map:* It is necessary to determine the nodes in the pedestrian network from where the walkers can get in the public transport services. The ideal situation is when the map data contains the street refuges connected to the public traffic network. If the refuges are not in map data, it is necessary to add the node representing refuge including the path connecting the refuge with the surrounding pedestrian network. The connection of the refuge into the pedestrian network is also important for the search of transfers between public traffic services.

*2) Mapping the Stops in Public Transport Network on the Street Refuges:* The network of public transport is typically created on the basis of routes of individual lines. Each stop in the itinerary of the line is identified by the name of the refuge where the service stops. If there are several refuges of the same name, there is a problem of how to create a unique mapping between the names of refuges in the timetable and the refuges in the map. This problem does not occur if the street refuges from timetables are identified by geographical coordinates. The carrier should know the position of refuges, where its services stop.

### G. Searching a Network of Public Transport

Paths search in the network of public transport is complicated by the fact that the value of each edge depends on the current time. Precise value of the edge is unknown until it is planned in some path.

*1) Unreliable Transfers:* If a transfer is realized within a single refuge, the following problem may occur. The timetable specifies only the expected time of departure from the refuge. In real traffic, there are deviations from the schedule. It has consequences especially for the line changing: If two services are scheduled at the same time at the same station, it is not possible to guarantee which service

arrives first in the real situation. If we consider the transfer between these services at the same minute, the transfer from the first service to the second one is possible, but the transfer in the opposite direction cannot be guaranteed. The timetable does not determine the order of arrival of the services.

*2) Length of Platform:* Until now, we were considering stop refuge as a point. However, the length of platform can be noticeable. If the passenger has to walk across a long platform, it can lead into several minutes of delay against the planned path.

It is appropriate to walk across the entire platform only in the case, when the passenger is getting in the service on the opposite side than he is getting off. By walking across the platform into the appropriate position, the passenger can spare some time. This can lead into a faster transfer and the passenger can catch earlier following connection.

The problem of passenger position at the platform makes sense only if the passenger arrives at the platform just in time of service departure. If the passenger comes earlier, then he is able to cross the platform into the appropriate position during his waiting for the service arrival. Likewise, if the passenger will be waiting for the following connection in the planned path, then the time needed to cross the platform after getting off the previous service could be subtracted from the waiting time.

## IV. Some Solutions

When implementing a prototype of a complex navigation for pedestrians, we applied some solutions to problems mentioned in section III.

### *A. Linking the network of walk paths and public transport network*

In order to move freely between search network of the public transport and search network of walk paths, it is necessary to connect both networks in certain nodes. The connecting nodes should be the street refuges, where the passengers are getting in and off the services of public transport.

*1) Street Refuges in the Map:* The street refuges were missing in the map base available to us. We get the positions of street refuges from other source. It was therefore necessary to correct the coordinates of refuges and it was necessary to connect the refuges into existing network of walk paths.

*2) Mapping Stops in the Network of Public Transport to the Street Refuges:* In our case, we did not have mapping of the street refuges of public transport to the places in map base. Nodes in the network of public transport are identified only by the stop names. For each stop, we had several refuges, representing different places in the map base. It was therefore necessary to distinguish the stops in the network of public transport, according to a service of public transport that is stopping at the given street refuge.

It was not possible to separate various street refuges of one stop in the public transport network. On the basis of practical experience, we know that the lines of public transport stop at different refuges. To be able to perform the mapping, we had to manually record a set of services that passes the given street refuge. So, we have assigned a line and direction to each refuge. However, this was not enough for unique mapping.

More complex situations may appear if one line is going through more refuges of one stop. We had to add a lookout for one stop forward and one stop backward on the line route. Still more complicated situations may appear where this approach will not work. The mapping created by matching refuges to stops requires maintenance in the case when the route of some line is changed. It is preferred that positions of the street refuges are identified directly in the data from the carrier.

### *B. Searching a Network of Walk Paths*

On the basis of map data, we had available, we created a search graph for the network of walk paths. The vertices of this graph are crossing or closing of some polyline. Specific nodes of this graph are the street refuges, which hold the identification of corresponding stop in the public transport network. So it is possible to move continuously from the walk paths search network to the search network of public transport.

Each polyline is represented by two oriented edges being to each other in opposite directions. The value of the edges determines the duration of walking, which usually depends on the walking speed and segment length. The problem occurs in sections with superelevation or some kind of barrier and at the crossings. The edge value may vary depending on the direction and can be even dynamic. In these cases, the details from map base are very important, because they determine the duration of walking in the given section. The duration should be parameterized by the actual dispositions of the user.

*1) Network Reduction:* When converting vector data to a network, it is possible to make a simple reduction of the vertices, where there is no branching of the graph.

*2) Entering Position on the Map:* When entering the starting and target position on the map, it is necessary to determine precisely the walk path that is closest to the user. We do not know the path from general position to the closest walk path, so we approximate it by a direct line. The selection of the closest walk path is given up to the user (see Figure 2). This choice can be complicated, and if the automatic approximation fails, the user can make a correction immediately according to his knowledge of the current location. In the worst case, the user will rely on the automatic choice.

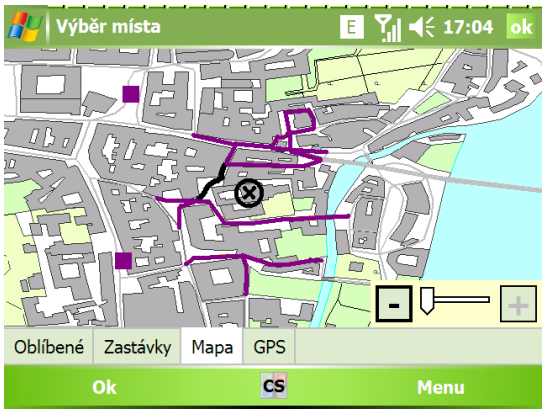



Figure 2. Starting position selection
The cross marks the selected starting position. The nearby walk paths are highlighted. The chosen section of walk path is the one closest to the starting position.

### C. Searching a Network of Public Transport

We have created the search graph for the network of public transport from available timetables. Vertices represent stop refuges, each refuge has its position on the map from where the user can continue in walk path.

Each edge represents a possibility to take a service to the next stop on the route of the line. The edges are characterized by a value, which determines the duration of travelling to the neighboring refuge. But if the user should get on a service in the planned path, then it is necessary to add a waiting time to the value of currently planned path. This time is derived from the current time (when the user gets to the stop according to the path plan) and valid timetable of the service he is waiting for.

Other approach could be representing every departure of a service as a vertex. This approach is robust for complex scenarios but not necessary in our case. Moreover this approach shows significantly lower performance as described in [12].

*Graph Reduction:* An interesting method of rail network reduction is described in [13]. Under certain conditions, it can be used to reduce the network of city public transport and so speed up the planning significantly. Fortunately, it is possible to solve the task in acceptable time [14]. Such reduction is generally an NP-hard problem [15]. Using approach based on this reduction the computational complexity will decrease to the level, where the path planning itself can be computed on portable devices in reasonable time [16].

The timetables determines the dynamic part of the graph. The representation of timetables should deal with the irregularities of a real world. One of the interesting approaches is described in [17].

*Unreliable Transfers:* If the planned path includes a transfer inside one refuge between two lines, which are leaving in the same minute, we are not able to ensure the order of services in most cases. We handle the situation by searching the departure time of the following service from the next minute from actual time. So, it cannot happen that we plan a transfer, which the user cannot make.

### D. Path Planning in Combined Network

If we consider the task in general case, we are given the starting and the target position on the map and we want to find a path to connect them. When planning, we start planning of walking routes from the starting position to all public transport stops in a particular area. Similarly, we plan walking routes for the target position. Walk paths should be searched in the opposite direction if, for example, super elevation has to be taken into account.

The general task of finding combined path is reduced to the task of planning connection in the public transport network. When planning transfer between services, it is still needed to use the walking graph to determine the transfer duration.

*Searching Walk Path to the Stop:* When searching for walking path from the starting position, all stop refuges in a given area are relevant to us. The passenger can use public transport service after reaching the refuge. The user can adjust the size of area, according to the distance he is willing to walk. Due to this limitation it is not necessary to search the entire network of walk paths, but only a relatively small part.

Due to the breadth first search, we are able to find all paths from the starting position to all street refuges in the area at once. The following search is made in the network of public transport, where the starting positions are the street refuges reached by walk, and their initial estimation of shortest path length is duration of the walk from the starting point.

*Precomputation of Transfers:* To avoid searching over both networks simultaneously, we performed precomputation of walk transfers and added special edges representing transfers into the network of public transport. As a result, we do not have to leave the search graph for public transportation during the search, so the overall branching of computation is decreased.

As it is a precomputation, it is necessary to specify the maximum length of walk transfer between services while creating search graphs. If we do not limit the length of walk transfer, it would lead to unbearable increment of branching of public transport search graph. It is unreliable for both computational and memory demands. On the other hand, if we choose the limit of walk transfers too strict, the path planning possibilities would be reduced. Some of the transfers would not get into the search graph due to precomputation. The limit of walk transfer is determined during the compilation of data, mostly based on experience and tests.

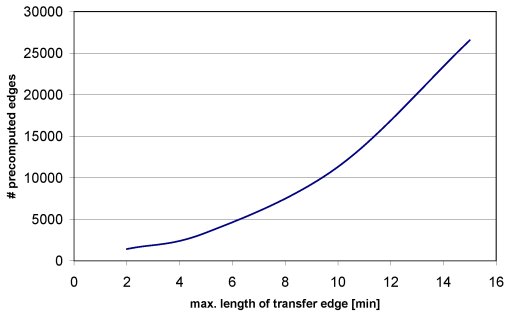



Figure 3. Precomputed Transfers Limit
The number of precomputed edges in dependency of the limit of maximum walk transfer distance.

The Figure 3 shows the growth of the number of precomputed edges with the increasing limit of walk transfers. For higher values, the number of edges in search network rises more than twice after adding the precomputed transfer edges.

*Searching Connection in Public Transport:* Due to previous steps, the planning of path is reduced into the searching path in the enriched graph of public transport. A specific part of this search is that in addition to previously found paths it is necessary to remember the current time. The waiting time for a service changing is determined according to the current time and timetable that is currently valid. On the basis of the current time other parameters of the dynamic network could be determined.

The user parameters and preferences should be taken into account when planning the path. In particular, the maximum number of transfers, the maximum length of walk section (the user is willing to walk continuously only for a certain distance), walk speed and more.

### E. User Preferences

Users may have very different movement dispositions. This can significantly affect planning of the path. We tried to take at least basic user parameters into account.

*1) Walk Speed:* All walk edges in search graphs are valued by the duration rather than length. Except special edges, this value is a walking time. The walking time is determined by the length of walk section, which is represented by the edge, and the default walking speed, for example, 5km/h. The appropriate correction of length of the walk section is performed only if the user changes the default walking speed.

Special edges are distinguished by additional indication of the character of the section, which is represented by the given edge. The value of special edges is not affected by walking speed.

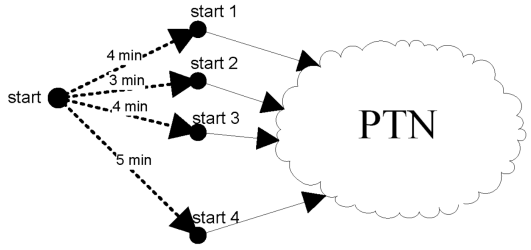



Figure 4. User Defined Starting Position
The user defined position is connected to the search network. The connection planning mechanism starts from all the assigned starting positions simultaneously counting in the initial path length estimation.

*2) The Maximum Length of Continuous Walking Section:* This parameter is added especially for precomputation of walking transfers, which are theoretically limiting possibilities of walk transfer. The parameter will lose effect if it is set to a value higher than the limit of precomputed walk transfers. At the same time, this parameter limits the size of the area for searching walk paths to the closest street refuges.

### F. User Places

In everyday use of our navigation, some set of places will be used frequently as starting or target points of path search. These places can represent home, work, school, etc. In these cases, most users already know the walking path, and know the time that it takes to the stops in certain area. Therefore, we have introduced a possibility to predefine these places, including duration of walk paths to the stops. Predefined positions reduce the time needed for user input and increase user comfort of the application.

### G. Reliability of the Network

A sudden reduction of the transport network may occur during the travel, for example, due to a technical fault of the route or vehicle. In this case, the current path plan may be irrelevant and needs to be recomputed according to the new situation.

*1) Excluding Line:* In our application, we allow the user to react to a situation similar to the exclusion of a particular line, which is affected by the failure. Any number of lines can be excluded from the search. After that, the path will be planned using other routes.

*2) Excluding Section:* Failure of a part of the network can affect both public transport and walk paths network. In both cases, it is necessary to allow the user to identify the affected part of the network and reschedule path plans another way. This action may require experienced user, and we do not solve it in our application.

## V. Reliability of the Connection in Public Transport Network

In this section, we will consider the reliability of the connection in public transport network as a probability that the service will not be delayed. With increasing value of the delay, the reliability will decrease.

The reliability of the connection found can be one of the user requirements on the path plan. But it can also be an additional information for planning the robust connection. For example, if we plan the path using not the fastest but reliable services it could be better than planning a slightly faster path using unreliable services. It is probable that the fast but unreliable path will fail, and the reliable path will be faster in the real situation.

The public transport service provider has often detailed information about the real movement of the vehicles. The differences between schedule and real situation are important for setting the path reliability.

### A. Real-time Information about the Delay

The ideal condition for planning the path to minimize the delay is to know the real position of the services. In that case, the path plan does not have to be based on the schedule, but can be directly assembled according to the actual situation in public transport network. The complication is that the situation can change during the realization of the path. It would be necessary to recompute the path plan dynamically in order to reflect the actual situation.

This solution puts high requirements on connection between public transport services provider and target application. It also requires a higher computation capacity to manage the dynamic path recomputation.

### B. Delay Dependency on a Daytime

The delay of the service can be caused by a periodically repeated event, for example, morning traffic jams. To detect these events, it is necessary to analyze the detailed data from the public transport network provider in a certain time range. Based on the analysis, the prediction of the delays can be propagated into the path planning mechanism. This analysis can be useful for the public transport provider as well.

The Figure 5 shows values of delay measured between two check points in public transport network for a single line in a three different days of week. The delays are measured according to time of a day. Negative values indicates that the delay is decreased in given pass of the section.

Furthermore, if we consider the path reliability as a frequency of services, we can recognize the dependency on the day time. The frequency of services is derived from the schedule, so it apparently has a periodical character.

### C. Delay Dependency on a Path Section

There is a question on where to count the delay prediction. It can be count for every combination of a stop refuge, service, and time of a day separately. This will lead to a large set of data. We can use the results of network reduction described in [16].

The principles of network reduction come from the similar

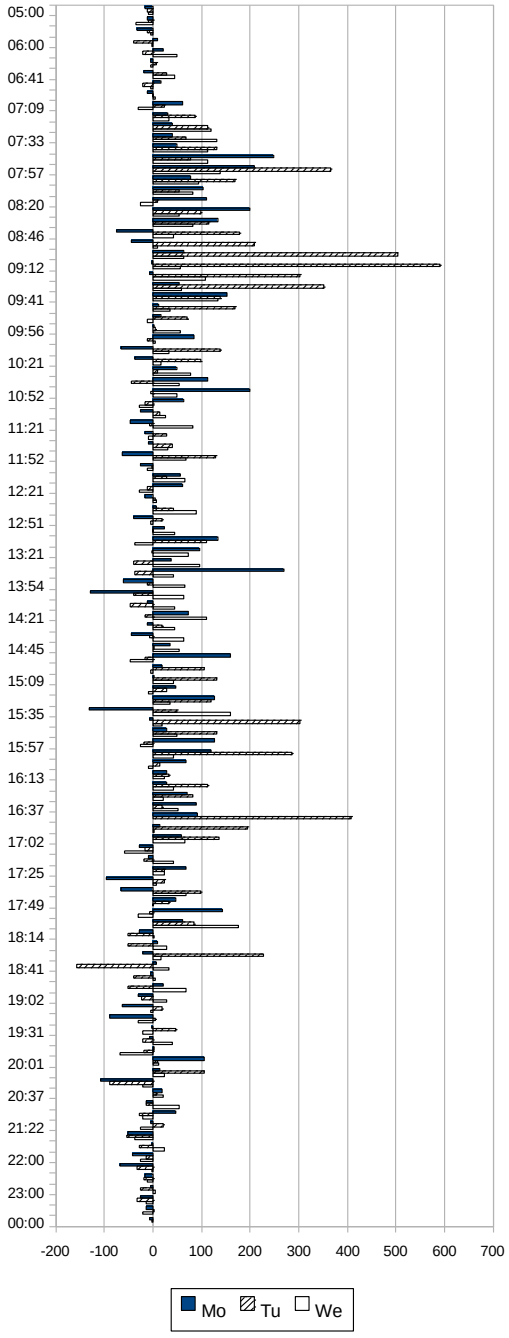



Figure 5. Section Delay for Various Days
Vertical axis shows the time of a day, horizontal axis shows the value of section delay in seconds.

behavior of several services in certain path section. This corresponds to the delay prediction. It is probable that the services running in a certain section will have similar results on the delay prediction. The events causing the delay would affect all the services same way in the certain section. This could reduce the problem to setting the delay prediction for whole path section.

Let us have a certain delay predicted for the given section in a given time of a day. It means that if the service arrives the section with some delay, it is probable that the value of the delay of a service after leaving the section will be increased by the value predicted for the section.

The Figure 6 shows values of delay measured between two check points in public transport network for three different lines.

It can be seen that in the early morning and late in the night, the delays are minimal. During the morning and in the afternoon the delays increase. The character of delay shows some similarities during the day among the services in the given section.

### *D. The View of Public Transport Service Provider and the Passenger*

The public service provider often watches every service instance separately. That means that the delays are counted in absolute value. In contrary, the passenger does not care about the certain instance of a service. He counts the delay of a service against the schedule.

For example, if the delay of some service is higher than the interval between two following services of the same designation, the passenger will count the delay against the previous scheduled service. In contrary for the service provider, the delay will be counted against the real scheduled instance of given service.

This situation can lead into misinformation of a passenger, which can think that the service arrives even sooner that it is scheduled. Nevertheless, the travel time will be increased in consequence of the delay.

## VI. USE CASES

### *A. Path Plan Dependency on Starting Time*

The following example of a path plan shows two paths between the same places. The only difference on the input of planning is the starting time. The first path starts only one single minute earlier. This situation shows, how critical is the current time and reliability of schedule of public transport services for the resulting path plan. See Figure 1.

In the case of failure in the network, the situation will be similar. The new recomputed path will have a significantly different plan in comparison with the original path plan leading through the unavailable part of network.

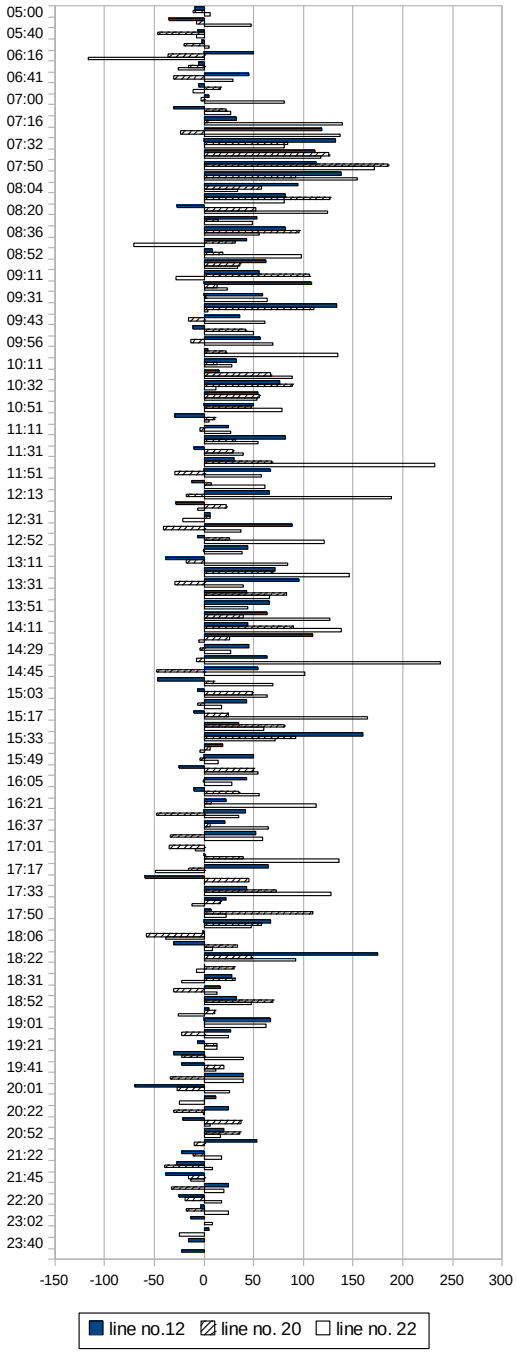



Figure 6. Section Delay for Various Services
Vertical axis shows the time of a day, horizontal axis shows the value of section delay in seconds.

### *B. Example of Advantageous Walk Transfer*

Benefits of combination of walk paths with the public transport will appear in situations where it is better to walk to a distant stop concerning the overall length of path.

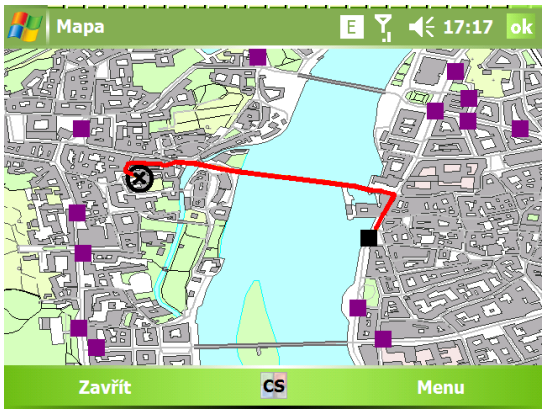



Figure 7. Advantageous Walk Transfer

For overall path length, it is advantageous to consider the possibilities of travelling from all stops in certain area. Sometimes the long initial walk section could lead to a shorter path, especially in town areas separated by river or other obstacles.

The Figure 7 shows initial walk segment of path plan leading from the starting position to a remote stop.

## VII. Practical Issues

When implementing real navigation system, the developers may face to many issues. Let us mention some of them.

### A. User Interface

A basic interface usable for most people can be described as graphic screen with maps and menu where users may use pointing device (touch screen, mouse, or joystick) and occasionally also enter some textual input. But what can we do if our users are unable to look at the map and follow the displayed hints?

One can propose voice navigation using hands-free. It is used in car navigation system, so it is a proven technology. For people able to watch the traffic well, it is usable to listen to the navigation system. When the visibility is worse, it can be useful (and for pedestrian moving in dark or fog or being blind it is quite common to behave so) to listen to the surrounding sounds very carefully. Even in the day, it can be important to know that some car is getting closer from behind. So, we must be very careful in using voice navigation. At least, the user must be able to state when it is possible to listen to the navigation and when to other sources.

### B. Data Precision

Most cars behave similarly and can use most of the paths in the same way. The problems may arose by large or very heavy vehicles. This issue can be solved when they occur (it is not possible to go through, let us go around) or in advance by extending the system by some additional information. Most limitations of the path are somehow indicated by special signs. The vehicles are divided into classes according to its characteristics. The characteristics of the class must satisfy the limitation of the selected path.

Also pedestrians may have different characteristics and movement limitations. Some of them cannot see. Others use wheelchair. Others have baby-coach. Others simply have problems with using stairs or very steep roads.

So, the limitations for pedestrian navigation may be binary (wheelchair are not able to go upstairs), sometimes it is not so strict and should be solved in a broader context (is it sometimes better to elevate the baby-coach a few stairs than go around for several minutes). Similarly, with baby-coach it is usually possible to use high-floor bus but it is less comfortable than going by (getting in and out) a low-floor vehicle. Then the hint must take into account how big time and price penalty the user is willing to pay for restricting himself to low-floor vehicle. The users must be able to specify their quite complex preferences and the system must be able to evaluate more possible paths according different restrictions.

Support for wheelchairs and baby-coaches require more precise data than necessary for general users. It must be ensured that the way is wide enough and smooth enough to be used by the users. The data have to contain such details like, for example, the height of the sidewalk above the road if we want to navigate user through a crosswalk.

### C. Data Cleansing

A question arise how to keep the data with all the detailed information actual. There are many subjects that can change the path properties. Some of the subjects announces the changes, so the data provider can propagate the change into the actual data set.

There is an option to keep the data up to date and furthermore to make the data more precise. We expect the navigation is running on a mobile device equipped with GPS receiver. Then it is technically possible to collect the real data as the user is moving along the path found by the navigation. The problem is that the user position is a kind of private information.

The user himself should directly decide if he wants to record his position while walking or not. The recorded data could be used locally. The application can make the path planning more accurate if the same path is overtaken repeatedly. The user himself should directly decide if he wants to share this recorded data with other users and/or the data provider.

The data collected by volunteers represent the real movement of certain types of users. This opens the whole area on how to utilize this data and how to verify their authenticity. This comes to another thought. The data provider does not have to be only a central authority. The collected data can be shared peer-to-peer by users itself. For example, the first

user that encounters an unexpected obstacle in the path can propagate it to the others. Then the navigations of other users can adapt the path plan to this obstacle. The problem how to verify the authenticity of an information about obstacle arises again.

### D. Data Timeliness

According to our experience, the original data from their owners are changing only rarely – correspondingly to the data owners' needs. The proper service of the navigation system requires data to be always up-to-date.

Another issue is that the users update their local copies of the navigation data only time to time. It is therefore required to mark the data with validity intervals (sometimes it is, for example, known how long a diversion will be valid).

Moreover, it can be necessary to adapt to changes faster than they arrive from the data owners – a data update mechanism. It is possible to make some updates using data measured by system maintenance team. Such solution has some limitations and brings additional costs. On the other hand usability of such system is then higher.

### E. Data Sources

When creating complex service, the data set could come from several different data sources. We have described the problems of combining the data for the search procedures and what kind of data can be needed for the final application. It is probable that the data are owned and/or managed by different authorities.

Several problems arises. First, the distributing responsibility for keeping the data updated. Second, the ownership of the data created by combining different search networks or the ownership of the changes in available data.

The availability of data is not given only by technical issues, but can also depend on the willingness of authorities managing and/or owning the data.

The usability of the final application depends on available data and on the ability to maintain the data. When concerning multiple different data sources and entities, which provide the data, a service-orientd architecture can be an advantageous solution. At least it will be advantageous for the part of application handling the data set preparation.

## VIII. System Structure

When creating complex navigation system, it is necessary to handle a large amount of data from various sources. Several tasks can be done repeatedly (for example, when preparing updates of navigation data.) It is advantageous to separate the task into several independent processes.

Reasonable complex navigation systems would share at least some of the requirements and can have similar structure.

### A. System Requirements

It could be advantageous (depending on the business model) to equip the system with both on-line and off-line access. The system should therefore have on-line access point, client software and data distribution subsystem.

Usually there are more sources (owners) of the needed data. We would need data integration.

The data from original sources may be distributed once upon a time. The system is expected to have the latest data possible. There must be an opportunity to incorporate changes in the environment that happen between the source data updates. A tight cooperation with traffic control centers can be an advantage.

The system should cover at least these parts:

- import data from their owners,
- data synchronization/integration,
- data management/maintenance (including information on changes),
- data modification to match navigation system needs,
- on-line navigation service,
- data distribution,
- client software for off-line navigation,
- interface for collection of data changes.

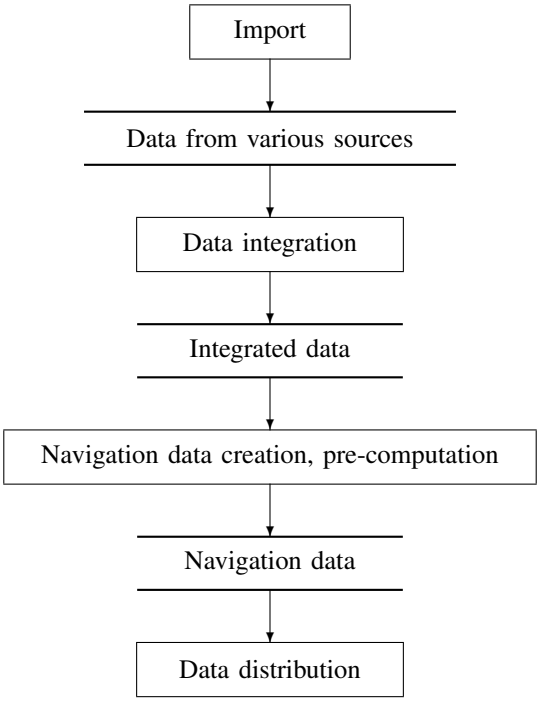



Figure 8. Simplified system structure

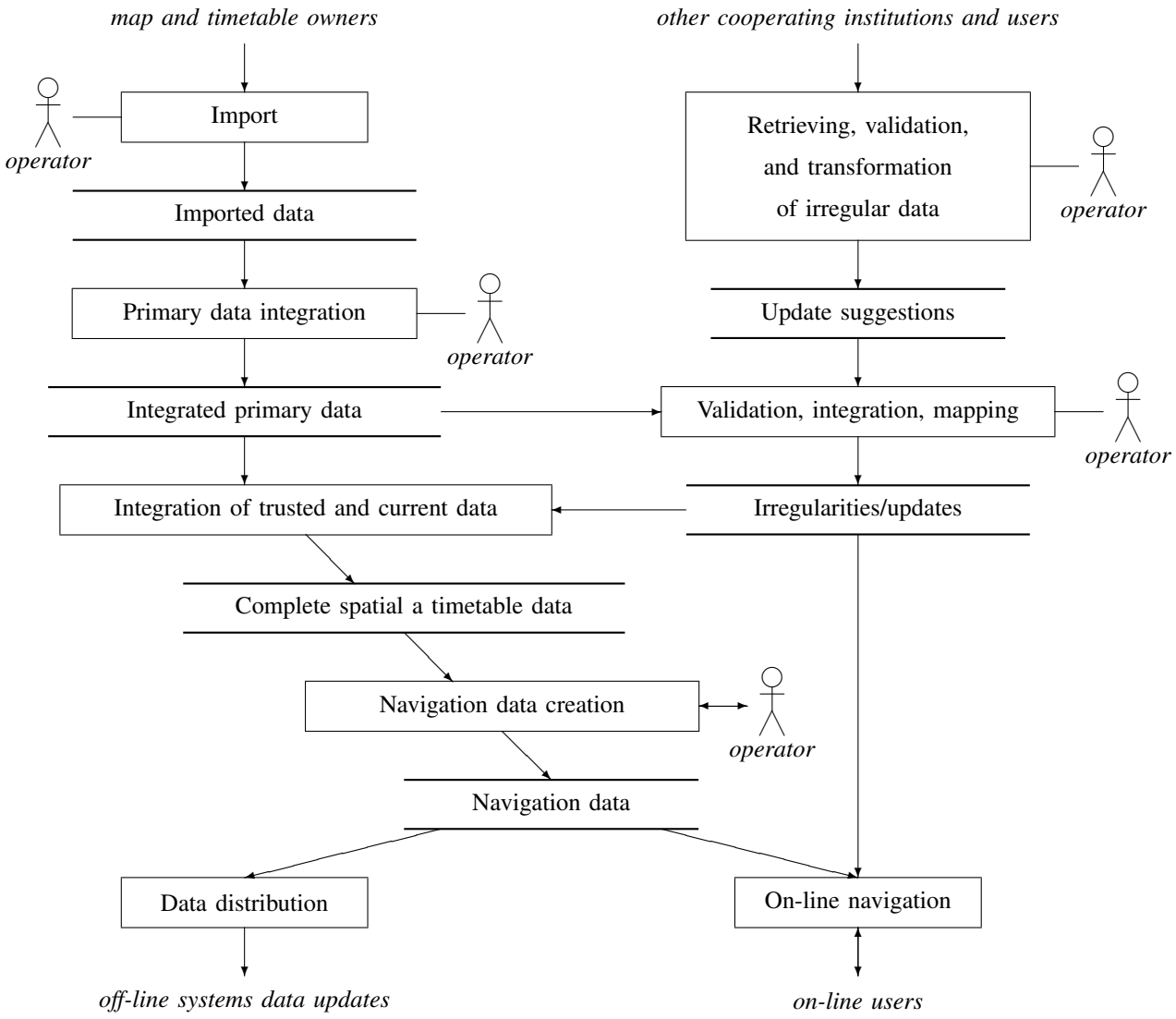



Figure 9. Navigation system architecture overview

### B. Necessary Parts

Considering overall design, there are parts necessary to ensure certain quality of provided navigation service. The Figure 8 shows the basic sequence of processes and data stores. The result of this sequence is a navigation service provided on-line or data sets for distribution to off-line clients.

The on-line service can take advantage from access to the actual data updates. So, it is possible to offer temporary or unverified data updates directly to the user. The Figure 9 shows the basic solution including the update mechanism.

### C. Update Sources

One of the problematic parts is the creation of data updates. The suggestion for the updates can come from different sources. In the case of an unreliable source of suggestion, the updated data should be verified before propagating to the end user. The other option is to mark the unverified data and leave the decision to the end user.

## IX. Future Plans

### A. Multicriterial Path Search

So far, we discussed only finding the time shortest path with some restrictive conditions. Requirements on the final path plan may vary and may not always be strict conditions. To be able to take into account various preferences it will be necessary to perform multicriterial search on a combination of networks. One of the promising approach, is to find Pareto-optimal solution for multicriterial path search, which is studied for railway networks in [18].

A computational complexity may exceed the possibilities of portable devices. It is therefore appropriate in the context of this approach to consider a different approach to the overall solution.

*1) Reliability of the Path Found:* One of the characteristic features of planning in public transport is the fact that the timetables are only a prescription for service scheduling. In real cases, the services can be delayed or cancelled. In the case, some of this situation is frequent, we can count on a certain probability that the service comes on time or will have a certain delay. If we know these probabilities, we can take the reliability of the connection into account when planning the path. Alternatively, if the user requires a reliable path, we can adjust the planning to handle the most probable delays.

Moreover, the following situation may arise. A service with less frequent intervals can occur in the path. Missing this service would mean a serious time loss for the user. In this case, it is appropriate to plan the route so that even in bad traffic conditions with high probable delays, it would be possible to guarantee a high probability of catching the critical service.

*2) Points of Interest:* Like in the case of ordinary tourist navigation, we should also be able to add points of interest. So the route plan could be adapted to the requirement to visit a point of interest or a category of points of interest, which is located closest to the direction of the planned path. These ideas are based on the assumption of multicriterial path planning.

## X. Conclusion

The combination of different types of navigation can bring advantages as well as disadvantages. Moreover, the combination of different networks can bring us into specific situations.

### *A. Advantages of Combination of Two Different Types of Navigation*

*1) Path Efficiency:* Combining the two networks gives us much more scheduling options than using only one type of navigation. Moreover, for walk sections we have far more information than if we use only the navigation in public transport. The resulting path plan does not have to estimate the transfers' duration, but is more accurate, because the transfer path is known. This allows us to plan more efficient and more reliable paths.

*2) Environmental Aspect:* We are trying to offer comfortable and accurate planning in the city using public transport to a wide range of users. This way we are increasing the comfort of the use of public transport and the level of transport-related services. The more users will prefer public transport over less ecological alternatives, the smaller will be the impact of urban transport on the environment.

### *B. Disadvantages of Combination of Two Different Types of Navigation*

*1) Different Planning:* We try to combine two very different networks. Each of them has different rules and heuristics, which can be successfully applied in one network, but may not be valid in the other one. It is therefore necessary to separate the search. On the other hand, the whole travel plan should meet the common criteria. To achieve this, it is often necessary to use a different mechanism in each network.

*2) Different Sources of Data:* For a network of public transport we need data of timetables and data of the positions of stops refuges. Walk network needs map data, including details needed for navigation of walkers. The application needs data from two different entities. In the case of commercial deployment of applications, the question "how to split the profit?" arises.

### *C. Available Data*

While developing our application, we had data for the city of Prague available. Map data provided to us "the Czech Office for Surveying and Mapping". Although the map data were not initially designed for the operation of navigation, we managed to adapt mechanisms working with them, so that our application was able to bring reasonable results.

The available data have shown that the operation of the application is not limited by memory or computing capabilities of portable devices. In addition, a limited connectivity is sufficient to keep the data updated. For most European cities, the search parameters should be comparable, excluding much larger cities like Paris, London, or Moscow.

Later on, we had data of the real movement of public transport services in the city of Prague. The comparison of real positions of services against the schedule brings us new pieces of knowledge and also new questions.

### *D. Real Life Consequences*

Using the JRGPS application, we learned that it can be reasonable to change slightly our habits: In some cases, it is better to go on foot instead of waiting for public transport and in some other cases, it can be advantageous to change entry or leaving stop.

## Acknowledgment

This paper was partially supported by the Czech Science Foundation by the grant number 201/09/0983 and by the Grant Agency of Charles University under project 157710.

## References

[1] V. Martinek and M. Zemlicka, "Some issues and solutions for complex navigation systems: Experience from the jrgps project," in *Systems (ICONS), 2010 Fifth International Conference on*, 2010, pp. 92 –98.

[2] L. de Marcos, R. Barchino, J.-M. Gutiérrez, J.-J. Martínez, S. Otón, F. Giner, and R. Buendía, "Ciceron-e: Interactive tourism for smes," in *Best Practices for the Knowledge Society. Knowledge, Learning, Development and Technology for All*, ser. Communications in Computer and Information Science, M. D. Lytras, P. Ordonez de Pablos, E. Damiani, D. Avison, A. Naeve, and D. G. Horner, Eds., vol. 49. Springer Berlin Heidelberg, 2009, pp. 420–429, url source last checked 20.1.2011. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-04757-2_45

[3] Google Maps Navigation, http://www.google.com/mobile/navigation/, url source last checked 20.1.2011.

[4] Navitime, http://www.navitime.com/, url source last checked 20.1.2011.

[5] Nokia Maps, http://maps.nokia.com/, url source last checked 20.1.2011.

[6] TomTom, http://www.tomtom.com/, url source last checked 20.1.2011.

[7] Navigon, http://www.navigon.com/, url source last checked 20.1.2011.

[8] Mio, http://www.mio.com/, url source last checked 20.1.2011.

[9] P. Sanders and D. Schultes, "Engineering fast route planning algorithms." in *WEA*, ser. Lecture Notes in Computer Science, C. Demetrescu, Ed., vol. 4525. Springer, 2007, pp. 23–36, url source last checked 20.1.2011. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-72845-0_2

[10] D. Delling, P. Sanders, D. Schultes, and D. Wagner, "Highway hierarchies star," in *9th DIMACS Implementation Challenge*, 2006.

[11] J. Koszelew, "Two methods of quasi-optimal routes generation in public transportation network," in *CISIM '08: Proceedings of the 2008 7th Computer Information Systems and Industrial Management Applications*. Washington, DC, USA: IEEE Computer Society, 2008, pp. 231–236, url source last checked 20.1.2011. [Online]. Available: http://dx.doi.org/10.1109/CISIM.2008.43

[12] E. Pyrga, F. Schulz, D. Wagner, and C. Zaroliagis, "Efficient models for timetable information in public transportation systems," *J. Exp. Algorithmics*, vol. 12, pp. 1–39, 2008, url source last checked 20.1.2011. [Online]. Available: http://doi.acm.org/10.1145/1227161.1227166

[13] K. Weihe, "Covering trains by stations or the power of data reduction," in *Proceedings of "Algorithms and Experiments" (ALEX98)*, R. Battiti and A. A. Bertossi, Eds., 1998, pp. 1–8.

[14] A. Liebers and K. Weihe, "Recognizing bundles in time table graphs - a structural approach," in *Algorithm Engineering*, ser. Lecture Notes in Computer Science, S. Näher and D. Wagner, Eds., vol. 1982. Springer, 2000, pp. 87–98, url source last checked 20.1.2011. [Online]. Available: http://link.springer.de/link/service/series/0558/bibs/1982/19820087.htm

[15] A. Liebers, D. Wagner, and K. Weihe, "On the hardness of recognizing bundles in time table graphs," in *WG*, ser. Lecture Notes in Computer Science, P. Widmayer, G. Neyer, and S. Eidenbenz, Eds., vol. 1665. Springer, 1999, pp. 325–337, url source last checked 20.1.2011. [Online]. Available: http://dx.doi.org/10.1007/3-540-46784-X_31

[16] V. Martínek and M. Žemlička, "Speeding up shortest path search in public transport networks," in *DATESO 2009*, K. Richta, J. Pokorný, and V. Snášel, Eds. Prague, Czech Republic: Czech Technical University in Prague, 2009, pp. 1–12, url source last checked 20.1.2011. [Online]. Available: http://ceur-ws.org/Vol-471/paper1.pdf

[17] R. Kasperovics, M. H. Böhlen, and J. Gamper, "Representing public transport schedules as repeating trips," in *TIME '08: Proceedings of the 2008 15th International Symposium on Temporal Representation and Reasoning*. Washington, DC, USA: IEEE Computer Society, 2008, pp. 54–58, url source last checked 20.1.2011. [Online]. Available: http://dx.doi.org/10.1109/TIME.2008.26

[18] M. Müller-Hannemann and K. Weihe, "On the cardinality of the pareto set in bicriteria shortest path problems." *Annals OR*, vol. 147, no. 1, pp. 269–286, 2006, url source last checked 20.1.2011. [Online]. Available: http://dx.doi.org/10.1007/s10479-006-0072-1

# UbiRoad: Semantic Middleware for Cooperative Traffic Systems and Services

Vagan Terziyan
MIT Department, University of Jyvaskyla
P.O. Box 35 (Agora), 40014
Jyvaskyla, Finland
e-mail: vagan@jyu.fi

Olena Kaykova, Dmytro Zhovtobryukh
Agora Center, University of Jyvaskyla
P.O. Box 35 (Agora), 40014
Jyvaskyla, Finland
e-mail: olena@cc.jyu.fi, dzhovto@cc.jyu.fi

*Abstract*—**Emerging traffic management systems and smart road environments are currently equipped with all necessary facilities to enable seamless mobile service provisioning to the users. However, advanced sensors and network architectures deployed within the traffic environment are insufficient to make mobile service provisioning autonomous and proactive, thus minimizing drivers' distraction during their presence in the environment. An ideal system should provide solutions to the following two interoperability problems: interoperability between the in-car and roadside devices produced and programmed by different vendors and/or providers, and the need for seamless and flexible collaboration (including discovery, coordination, conflict resolution and negotiation) amongst the smart road devices and services. To tackle these problems, in this paper we propose UbiRoad middleware intending utilization of semantic languages and semantic technologies for declarative specification of devices' and services' behavior, application of software agents as engines executing those specifications, and establishment of common ontologies to facilitate and govern seamless interoperation of devices, services, remote systems and humans.**

***Keywords- context-aware services; cooperative traffic; smart road; middleware; semantic technologies; agents***

## I. INTRODUCTION

There is about half of a billion drivers only in Europe, who wish driving to be more comfortable, efficient, ecological and less risky. Not far are the times when cars will themselves prevent accidents. People spend more time in vehicles and they are expecting also more possibilities to work and use various services while traveling, which requires new travel infrastructure and automation services [2]. These should combine various vehicles, their drivers and passengers, smart roads and appropriate Web services [3]. Recent wireless and internet technologies enable completely new possibilities to integrate available efforts into the new advanced traffic paradigm – cooperative traffic [4].

Service-oriented architectures related to traffic management, smart roads and future context-aware services for drivers are closely integrated into the Internet of Things [5], which is a world where things can automatically communicate to computers and each other, providing services for human benefits. In such "Future Internet", intelligence and knowledge will be distributed among an extremely large number of heterogeneous entities: sensors, actuators, devices, cars, road infrastructures, software applications, Web services, humans, and others. To realize this vision, there is a need for an open architecture, which will offer seamless connectivity and interworking between these heterogeneous entities. Moreover, ensuring collaboration, synchronization but also control of this distributed intelligence is a challenge that needs to be addressed, or the Internet of Things will become a chaotic, un-controlled and possibly dangerous environment since some actors of this Internet have impact on the real world (e.g., software or humans through actuators). Cooperative traffic domain enables interoperability between a large number of heterogeneous entities, while ensuring predictability and safety of their operation, is difficult without an extra layer of intelligence that will ensure the orchestration of these various actors according to well-defined goals, taking into account changing constraints, business objectives or regulations. This paper introduces such a middleware layer (UbiRoad). It provides cross-layer communication services (data-level interoperability) to the entities and extended multi-agent technologies will provide collaboration-support services (functional protocol-level interoperability and coordination) for these entities. The UbiRoad middleware concept apparently entails a vision of a multifaceted, multi-purpose and multipronged middleware platform applying multidisciplinary approach to extension and enhancement of the future smart traffic environments UbiRoad middleware should be rather seen as a meta-structure on top of the future intelligent transportation systems and services and as intelligent stratum between the smart road device layer and the future service oriented architectures.

A first major problem to be addressed by UbiRoad is inherent *heterogeneity*, with respect to the nature of components, standards, data formats, protocols, etc., which creates significant obstacles for interoperability among the components of ubiquitous computing systems. This heterogeneity is likely to induce some integration costs that will become prohibitive at a very large scale preventing a rich ecosystem of applications to emerge. It is generally recognized that achieving the interoperability by imposing some rigid standards and making everyone comply could not be a case in open ubiquitous environments. Therefore, the interoperability requires existence of some middleware to act as the glue joining heterogeneous components together.

The second major issue is to guarantee high level of *safety*. Since the IT infrastructure and through them users are going to have real actions in the real physical world through

various actuators we have to ensure that these actions are properly controlled and coordinated. Despite the wish to enable as many actors as possible to have access to physical world objects around the world to enable a large set of diverse applications, this should be done in a well-understood and safe manner. The "things" will have to exhibit some required behaviors that humans have adopted to assemble in cooperative traffic social interactions.

The UbiRoad approach can be seen as studying the triangle of device-software-human interaction seen from the perspective of the above described scenarios. Henceforth we refer to "device" as to any monitored or controlled physical objects including e.g., vehicles. Substantial research results related to edges and vertices of this triangle have been (recently) reported [6, 7, 8] (e.g., efforts related to middleware for embedded systems, efforts related to integration of diverse software systems and services, etc). What is missing is an integrated coherent approach to cover the whole triangle. Moreover, many on the past research initiatives do not truly deal with the core topic, which is *interoperability versus just interconnectivity*. The components of cooperative traffic systems should be able not only to communicate and exchange data, but also to flexibly coordinate with each other, discover and use each other, learn about the location, status and capabilities of each other, and jointly engage in different traffic situations. Moreover, the components must achieve the above using an always-on, safe, robust and scalable means of interaction.

Further in this paper, we argue in favor of fully interoperable (though heterogeneous), highly dynamic and extensible smart road environments. We present a specialized agent-driven middleware platform UbiRoad, in which each ubiquitous smart device (as well as each individual service exposed as an individually accessible entity through the environment) will be assigned a representative agent within UbiRoad. The resulting multi-agent system will be exploited as a mediation facility enabling rich cooperation capabilities (e.g., discovery, coordination, adaptability, and negotiation) amongst the devices inhabiting the smart traffic environment. Utilization of semantic technologies [9] in UbiRoad will ensure efficient and autonomous coordination among UbiRoad agents and will thus ensure interoperability between associated devices and services. Several UbiRoad ontologies are an important asset contributing to interoperability realization within future smart traffic environments. These ontologies are used not only for the benefit of UbiRoad middleware architecture, but also and most importantly for facilitation of interoperability and integration of existing and brand-new future devices, services and methodologies. Through appropriate declarative specification of smart road components' behavior and using sophisticated choreographic control agents in a multimodal dynamic networked environment, the UbiRoad enables various devices and services to automatically discover each other and to configure complex services functionally composed of the individual services' and devices' functionalities.

The rest of the paper is organized as follows: in Chapter II we are providing the motivating scenario for the new challenging requirements to traffic management systems; in Chapter III we list the requirements and related challenges to be addressed when designing such systems; in Chapter IV we provide possible solution for the challenges based on the concept and architecture of the so called Global Understanding Environment; in Chapter V we discuss some important and challenging features of appropriate agent-driven platform (UBIWARE) suitable for UbiRoad implementation, such as: semantic adapters and integrators (called *OntoNuts*); semantic visualization technology (called *for-eye*); user-driven system configuration (via so called *smart comments*); and *semantic blogging*; in Chapter VI we overview appropriate software architecture; in Chapter VII we briefly discuss on Traffic and Mobility ontology and system integration; in Chapter VIII we briefly comment on related work; and we conclude in Chapter IX.

## II. MOTIVATING SCENARIO

Consider the following story, in which we try to integrate several possible scenarios of future use for the UbiRoad middleware.

(*Beginning of the story*) "Timo lives in Jyväskylä. Former researcher, he is a widely recognized expert in the field of intelligent software agents. Nowadays Timo owns a small IT business based in Jyväskylä, and his firm is often subcontracted by large IT and telecom enterprises to perform highly specialized development services. Therefore, Timo is a frequent guest in Helsinki and Helsinki region, where most of his company's employers reside. Despite considerable distance between Jyväskylä and Helsinki, Timo likes neither airplanes, nor trains, and always travels inside Finland by car. Fortunately, he is a big cars-lover and a good driver.

Timo arrived in Helsinki early in the morning and spent the whole day participating in a few various business meetings and research seminars. Now, when he is about to leave Helsinki, he feels very tired. He could stay in Helsinki overnight, but he has another important meeting scheduled for tomorrow at 7 am in Jyväskylä. Not to fall asleep on the way back to Jyväskylä, Timo drops by the nearest cafeteria and drinks a cup of strong coffee. Having felt a burst of energy after the sprightly drink, he gets into his car and leaves Helsinki at dusk. In the car Timo selects from his audio collection some nice invigorative music to listen to and sets the car control system's operating mode to 'exhaustion' using the available on-board control panel. Timo knows that in this mode the awareness levels of a multitude of software agents, which inhabit his car and make it a part of the UbiRoad intelligent transportation system, reach the highest possible value. Now he feels much less vulnerable because of fatigue, as agents in this mode help significantly reduce exposure to various on-road risks. The in-car control system adjusts climate conditions (temperature, humidity, level of oxygen) to optimal levels with respect to the selected operating mode: it aims to maintain cool, fresh, oxygen-rich atmosphere inside the car in order to prevent the driver from falling asleep; and activates on-board alarm system, which is configured to give to the driver light and audio indication

every 30 seconds (not allowing him falling asleep). The corresponding UbiRoad traffic agent (representing Timo and his car as a dynamic road user entity) sets own hazard level to 'red', thus notifying other road users of potential risks associated with its road user. When Timo drives onto the motorway going out of Helsinki, he feels much more comfortable and relaxed, as he is sure that all necessary measures of passive risk prevention have been undertaken and as he should no longer pay attention to the oncoming traffic (on motorways directions of traffic are separated).

To shake himself up a bit, Timo switches to the left lane (each direction of the motorway connecting Helsinki and Lahti has two lanes) and starts overtaking all cars, which slowly go on the right lane. The speed limit on this motorway is 120 kmph and driving at it can be refreshing. After some time of such racing Timo however forgets about the speed, which immediately goes beyond 130 kmph, and the traffic agent monitoring velocity and controlling speed regimes detects inadmissibly excessive speed (via comparing actually measured vehicle's velocity with the speed limit that the agent can read from RFID-enhanced traffic signs located on the sides of the road) and activates a loud beep tone combined with an appropriately marked blinking red LED on the control panel. Timo takes this as a timely signal to calm down, decelerates to the allowed speed limit and uses the cruise control functionality embedded into the steering wheel to fix the speed at the current level. Now he can release the accelerator completely and give his leg some rest.

Timo utilizes voice control system to engage a travel estimator service and thus to find out the approximate time of his arrival in Jyväskylä. The specialized voice recognition system reads Timo's oral instructions, interprets them and finally transforms them into the format recognizable by UbiRoad agents. Then the corresponding communication agent finds an appropriate travel estimator service in the Internet, negotiates service contract with the agent representing the service and finally invokes the service. The result of travel duration estimation, 2 hours and 10 minutes, appears on the LCD screen built in the control panel to the right of the driver's seat. Timo decides to call home and let his wife know he is coming back soon. Timo's mobile phone is already connected with the in-car control system via Bluetooth. Timo utilizes voice control to access his phone and then voice dial to call Anna. While the picked number is being dialed, playing music is automatically damped down, and as soon as the phone connection is established, the conversation is output through the in-car embedded speaker system. After several minutes of chatting with Anna, Timo notices that he is driving already in the neighborhood of Lahti. Here 3G telecommunication network is available. The communication agent immediately detects this and using the LCD screen asks Timo if he is willing to switch to a video call. Timo accepts the offer by pressing the corresponding button on the touch-sensitive screen. The communication agent immediately requests the video capture service from a tiny camera embedded in the control panel in front of the driver's seat. Then it rearranges the current voice call session as a new video call session without interrupting the call and interweaves the audio component acquired through Timo's hands-free microphone with the video component obtained by the in-car embedded video camera. A live view of Anna appears on the LCD screen of the control panel. However, as shifting driver's focus to this side screen is inconvenient and distracting the driver from actual driving, the picture on the screen is instantly projected on the internal surface of the car's windscreen just in front of the driver's seat. The projected image is however semi-transparent not to impede driver's clear view of the road.

Timo finishes talking with his wife when Lahti is already left behind. He notices that twilight almost gave the place to solid night, but the motorway is still well illuminated. Timo decides to make a short stop at the picturesque roadside restaurant "Tähtihovi" in order to stretch his legs and have another cup of coffee before proceeding to the most difficult and boring part of his trip. Soon after this stop Timo should drive off the motorway to the side route leading to Jyväskylä. The traffic agent recognizes this major route change and reminds Timo of it well in advance using available visual indication means (LCD screen, projection on the windscreen, etc.) As Timo turns to the needed side road, he soon finds himself completely benighted as roadside lamps are uncommon here. He switches to upper beam to see at least something. Using embedded luminosity sensors, the agent monitoring external physical environment immediately detects severe lack of light on the road and activates built-in night vision system that multiply amplifies luminosity of the reflected light both in visible and infrared spectrum, thus being able to identify distant objects also by the heat they emit (e.g., oncoming cars, cyclists, pedestrians, elks, etc.). Such enhanced view of the road environment is projected on the internal surface of the car's windscreen so that it maximally coincides with the driver's field of view. Hence, Timo is now able to see everything much more clearly and recognize moving objects well in advance. What is more, in observed conditions of dark driving on a narrow bidirectional road the traffic agent starts to provide necessary assistance services such as improved navigation and automated signaling, e.g., a dynamically changing light-modulated traffic map of the neighborhood (specifically highlighting the route undertaken) is projected on the right side of the windscreen; upcoming turns and bends of the road are visually indicated (e.g., in the form of light arrows in the upper part of the windscreen); crossroads and cars approaching from the opposite direction are also identified for the driver in good time; switching from upper to lower beam (in proximity of oncoming cars) and back is performed automatically.

Luckily, the road is almost empty at night, and Timo almost reaches Jyväskylä when he catches up a heavy truck slowly going ahead of his car. Road is constantly dodging and the road-bed is narrow to comfortably overtake the truck. Timo almost loses patience waiting for a more or less

straight section of the road, and as soon as such section appears ahead, he confidently sends the car on the opposite lane and starts overtaking the truck. Suddenly he sees an opportune notification of an oncoming vehicle, which is still on the other side of the hill ahead of Timo and is thus unseen, but is quickly approaching. Perhaps, Timo is too tired as he makes an estimation error: he decides that he has enough space and time to complete the maneuver and continues overtaking. The oncoming car is however approaching too fast making head-on meeting with Timo's car almost inevitable. Moreover, the truck being overtaken turns out to be a long road-train, and it is already too late to get back behind it because Timo's car has passed more than a half of the truck's length already, when Timo realizes that he fell a victim to his own fatigue and impatience, and that only a miracle can now save him from head-on collision with the other car. UbiRoad intelligence is such a miracle.

The UbiRoad traffic agent that resides in Timo's car establishes communication with the approaching car's traffic agent immediately after it recognizes the presence of another vehicle in the proximity. At the same time it maintains communication with the traffic agent of the truck. The agents jointly monitor the process of rapprochement of the (three) vehicles. When Timo starts his overtaking maneuver, the traffic agents realize the situation is no longer standard. They integrate their individual traffic information, jointly reason upon it in the dynamic traffic context, and deduce that the collision is unavoidable. To prevent the traffic accident or any other dire consequences of Timo's mistake, the agents must undertake active measures of risk mitigation. The traffic agents of the approaching car and the truck notify their drivers of the potentially critical hazardous traffic situation and forcibly decelerate their vehicles to buy Timo enough time for successful completion of the overtaking maneuver. For its part, Timo's traffic agent aggressively visualizes the imperative "complete the maneuver", thus granting some extra confidence to its driver, who is already close to panic. Given such clear instruction, Timo accelerates even more and safely completes the overtaking maneuver. In twenty minutes, when he, exhausted as a squeezed lemon, but happy to escape probably fatal traffic accident, parks his car in his parking slot, another in-car agent reads Timo's schedule for tomorrow (stored in the organizer application within Timo's mobile phone) and sets engine warming-up timer to 6.30 am …" (*end of story*).

To be able to make this scenario a reality we have to face several challenges described in the next chapter.

## III. UBIROAD MIDDLEWARE CHALLENGES

### A. *Interoperability*

By proclaiming interoperability as its major ultimate objective, UbiRoad approach deals with three major types of interoperability problem: technical interoperability (being the capability of devices, protocols and other technical standards to co-exist and interoperate), semantic interoperability (being the capability of various system components to treat and interpret exchanged data and information identically and share a common understanding of it), and pragmatic interoperability (being the capability of system components to capture willingness of partners to collaborate or, more generally, to capture their (and even human users') intent). Technical interoperability will be achieved through the agent-based mediation between different devices and standards with the aid of special adapter components and tunneling mechanisms. Semantic interoperability is the main focus of the UbiRoad approach as it is a prerequisite for seamless information internetworking and integration, and for smooth autonomous communication between various resources within a smart traffic environment. Semantic interoperability can be achieved by exploitation of rich metadata describing informational objects and semantic resource descriptions written in compliance with well-established semantic standards and on the base of predefined domain ontologies and UbiRoad Ontologies. Pragmatic interoperability amongst smart space components is achieved through appropriate design of declarative specifications of such components' behavior and on-the-fly agent-based identification of this behavior using given descriptions. Finally, the most innovative type of interoperability, which UbiRoad provides, is the so-called 'cross-layer' interoperability, e.g., interoperability between devices and services in a smart traffic environment. This particular class of interoperability problems is often difficult to solve even on individual basis. However, UbiRoad provides native support for cross-layer interoperation by implementing the paradigm of resource-oriented networking. This paradigm enforces unified treatment of various system components, e.g., devices, services, applications and even users, as different types of resources (Figure 1).


Figure 1. Agent-driven smart road interoperability

The communication is then established between resources regardless their particular type provided that negotiation is performed by resources' representing agents

(associated with resources within smart traffic environments and beyond) as shown in Figure 1 and appropriate Semantic Web standards for unified resource description are used.

### B. Flexible Coordination

As smart traffic environments are basically deployed to provide users with dynamically configured, customized, value-added and on-the-move autonomously operating services, UbiRoad targets establishment of such service creation and provisioning framework that would emphasize the above mentioned characteristics of ubiquitous services. Customization, personalization, added value, dynamicity and autonomy of services is to be achieved through construction and utilization of context-aware, adaptable and reconfigurable composite service networks. Service networks can be composed using declarative specifications of service models. Reconfigurability of service networks is made possible via utilization of hierarchical modeling of service control and its run-time execution. Dynamic adaptation of services is performed by special context-aware control components built in service networks. The traditional tradeoff "customization vs. autonomy" can be dealt with through a balanced use of user-aware goal-driven on-demand service composition, AI-enriched active context-awareness capturing user intent, and user-collaborative passive context-aware service composition. Though it is a challenging task, utilization of agent-based approach for service composition makes it much more flexible compared to traditional orchestration approaches. This difference in flexibility can be seen from the definition of the traditional Semantic Web services (SWS) given in [18] ("Self-contained, self-described, semantically marked-up software resources that can be published, discovered, composed and executed across the Web in a task-driven way") and the definition of proactive (agent-driven) SWS given in [19] ("Self-contained, self-described, semantically marked-up *proactive* software resources that can be published, discovered, composed and executed across the Web in a task-driven way, *and which behave to increase their utility and are the subject of negotiation and trade*"). Agents can bring many valuable features into a service composition framework, e.g., precomposition, distributed hierarchical control of service networks (not requiring a dedicated underlying infrastructure), and enhanced negotiation of non-functional service parameters.

### C. Self-Management

UbiRoad brings self-management aboard via presenting totally distributed agent-driven proactive management system. UbiRoad agents monitor various components, resources and properties within the system architecture and infrastructures belonging or otherwise interacting with the managed smart road environment, and react to changes occurred by reconfiguring the architecture in appropriate way with respect to the predefined (or inferred) configuration plan. Configuration plans basically represent enhanced business models, which are adhered to during accomplishment of communication procedures between different parties. Due to purely distributed layout of the agent system and outstanding agents' programmability, merely all kinds of business models can be formalized and enacted by the UbiRoad management platform (due to richness of the utilized agent communication language and of the associated ontology base). In addition to this, UbiRoad agents are capable of learning via utilizing available data mining algorithms and further dynamically reconfiguring the managed architecture on the basis of acquired knowledge, thus being capable of inferring (also collaboratively) new configuration plans. UbiRoad can be deployed on top of any architectural model (including ad-hoc and peer-to-peer, which is of crucial importance for highly dynamic traffic environments) due to benefits of agent technologies and open resource interfaces. Also, the UbiRoad platform can make use of contextual information extracted from the managed networking environment in order to act as appropriately to the observed requirements and circumstances as possible.

### D. Trust and Reputation

Trust is identified as one of the major and most crucial challenges of the future computing and communications. We envisage a semantic ontology-based approach to building a universal trust management system. To make trust descriptions interpretable and processable by autonomous trust management procedures and modules, trust data should be given explicit meaning via semantic annotation. Semantic trust concepts and properties will be utilized and interpreted using common trust ontologies. This approach to trust modeling is especially flexible because it allows for various trust models to be utilized throughout the system seamlessly at the same time. Trust information can be incorporated as part of semantic resource descriptions and stored in dedicated places within the UbiRoad platform. Communication and retrieval of trust information will be accomplished through corresponding agent-to-agent communication. Agents representing communicating resources must be configured appropriately to handle all necessary trust management activities between the corresponding communication parties. Trust management procedures can be realized as a set of specific business scenarios in the form of agent configuration plans.

### E. Other Challenges

Specifically, due to utilization of extended intelligent agent technology UbiRoad significantly contributes to realization or enhancement of the following important characteristics and functionalities of collaborative traffic environments:

- Data mining and knowledge discovery (e.g., utilization of accumulated statistics of traffic accidents), which may be organized either by establishing centralized Web server with appropriate data processing services or by local processing of the analytics and exchanging of it in a P2P manner;
- Learning (e.g., case-based learning, when traffic agents can learn on sets of predefined examples of traffic situations);

- Global data and knowledge reuse (e.g., traffic environments have a common infrastructure, which inter alia provides means for storing and sharing of traffic information; agents may access external information sources located, for example, in the Internet);
- Enhanced traffic services (e.g., traffic services such as, for instance, traffic signs are RFID-annotated, which allows agents to identify them and conveniently communicate their meaning to drivers or take on appropriate actions);
- Enhanced collaboration between various road-users (e.g., collaboration between drivers on the road can be significantly enhanced and automated by the dialog between their representative agents; proximity-driven collaboration);
- Global context awareness, contextual filtering and visualization (e.g., traffic situations are treated by traffic agents in context (set of relevant contexts); traffic information can be displayed for a driver with respect to the observed context or as the reaction to contextual changes occurred; combined utilization of active context awareness (e.g., in critical situations) and passive context awareness (e.g., when user decision is required));
- Critical situation management (e.g., protocol-based collaboration, i.e., when agents recognize critical traffic situations, they can use corresponding predefined action plans for effective prevention/avoidance of these situations).

## IV. THE SOLUTION BASED ON GLOBAL UNDERSTANDING ENVIRONMENT

The solution for UbiRoad challenges is based on results of SmartResource [8] and UBIWARE [10] projects. Their objectives were research and development of the large-scale environment for integration of smart devices, web services and humans based on Semantic Web and agent technologies. The projects belong to the Industrial Ontologies Group [17] research roadmap towards the Global Understanding Environment (GUN) [11, 12]. When applying Semantic Web in the domains of ubiquitous computing and smart spaces, it should be obvious that Semantic Web has to be able to describe resources not only as passive functional or non-functional entities, but also to describe their behavior (proactivity, communication, and coordination). In this sense, the word "global" in GUN has a double meaning. First, it implies that resources are able to communicate and cooperate globally, i.e., across the whole organization and beyond. Second, it implies a "global understanding". This means that a resource A can understand all of (1) the properties and the state of a resource B, (2) the potential and actual behaviors of B, and (3) the business processes, in which A and B, and maybe other resources, are jointly involved.

According to GUN, resources (e.g., devices, humans, software components, etc.) can be linked to the Semantic Web-based environment via adapters (or interfaces), which include (if necessary) sensors with digital output, data structuring (e.g., XML) and semantic adapter components (e.g., XML to RDF). Agents are assumed to be assigned to each resource and are able to monitor data coming from the adapter about states of the resource, decide if more deep diagnostics of the state is needed, discover other agents in the environment, which represent "decision makers" and exchange information (agent-to-agent communication with semantically enriched content language) to get diagnoses and decide if any maintenance action is needed. Implementation of agent technologies and Multi-Agent Systems (MAS) within GUN framework allows mobility of service components between various platforms, decentralized service discovery, FIPA communication protocols utilization, and MAS-like integration/composition of services.

Agent-based layer of GUN-based architectures (e.g., of UbiRoad middleware), in addition to the agents, which are the representatives of the resources of interest, includes also an agent managing the *repository of roles and scenarios* encoded in RDF-based Semantic Agent Programming Language (S-APL) [13], an agent managing the *repository of reusable atomic behaviors* (i.e., software components that agents can load if a scenario prescribes), and an agent managing the *directory* that facilitates flexible discovery of agents (and thus of corresponding resources). S-APL – is a hybrid of semantics (metadata/ontologies/rules) specification languages, semantic reasoners, and agent programming languages. It integrates the semantic description of domain resources with the semantic prescription of individual and collaborative agents' behaviors.

### A. Universal Adapters for UbiRoad

Semantic adapters layer is one of the most important layers of GUN and UbiRoad architecture (see Figure 2). Ideally the adapter should be that kind of software that is able to automatically reconfigure itself for each new resource based on its declarative description. As a result of adaptation any parameters observed, measured or collected elsewhere about the resource will be available in the same semantically rich format (RDF-based) referring some shared ontology. We developed RscDF (Resource State/Condition Description Framework) as a subset of S-APL and an appropriate format for adapters output [14]. It extends RDF by making it more suitable for semantic annotation of dynamic and context-sensitive data about the resources. It provides opportunity to put any RDF statement into context, which is described by a container of RDF statements. Appropriate schema also includes some specific properties able to describe dynamic and if needed multilayered context of statements. In [20] it is argued that there must be at least four categories of ontologies to represent and capture context-sensitive sensor data (devices ontology – to recognize different devices in the environment; context ontology – to model environmental

information; data ontology – to make uniform of data coming from different sensors; and domain ontology – to represent a specific domain, e.g., collaborative traffic).

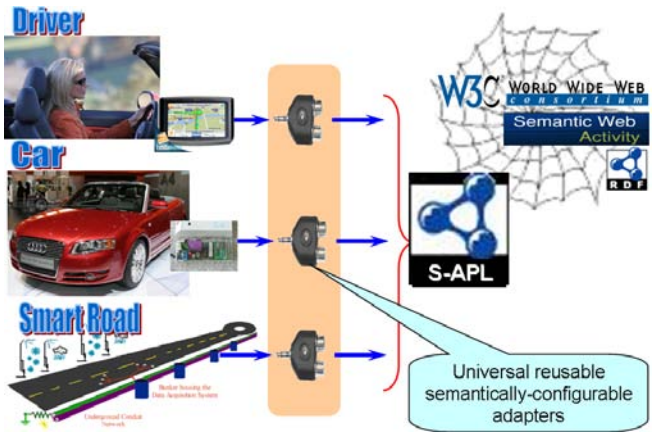



Figure 2. Semantic adapters for heterogeneous resources

## B. *Reusable Behaviors for UbiRoad Components*

Behavioral layer is another important layer of GUN and UbiRoad and it is designed to make every domain resource proactive, which means able to autonomously behave towards achieving certain goals depending on its role in the domain. Such behavior depends on the nature and the type of the resource, its placement in the environment, relations with other resources, environmental parameters, etc. In UbiRoad, autonomy and proactivity of resources were implemented by means of software agents. The main challenge however was to avoid designing different agents for each of heterogeneous resources but implement just one universal behavior engine for an agent (to make it like an "artist"), which will be able to play any declaratively described behavior according to its current role. We require designing such reusable declarative behavior descriptions to be made with as minimal effort as possible and with maximal reuse of previously designed behaviors and their components when designing new ones (see Figure 3). Ideally the agent should be that kind of software that is able to automatically reconfigure itself for each new resource based on declarative description of this resource role in the domain or within some business process. We designed RgbDF (Resource Goal/Behavior Description Framework) as S-APL subset and a tool for semantic annotation of behavioral properties of the resource (goals, plans, roles, actions, intensions, etc.). It extends RDF by making it more suitable for semantic annotation of data about proactive and autonomous behavior of the resources [15]. The extension (in addition to the features provided by RDF, OWL and relevant reasoners) allows making explicit links from behavioral properties of proactive resources to appropriate atomic software components, which are intended to implement described behavior when appropriate. The roles (i.e., appropriate behaviors) of agents can be chosen and changed depending on current context of the situation, and this means that each agent should be able to download from some shared place the description of a new role whenever needed. Taking into account that some situations may be time-critical to react and that available semantic reasoners are not always able to provide decisions in real time (see analysis made in [21]), the UbiRoad solution allows (when appropriate) combining semantic reasoning (e.g., automated online generation of the actions plan in S-APL and implementing it) with the plans compiled in advance and available as hard-written (e.g., in Java) reusable atomic behaviors [13].

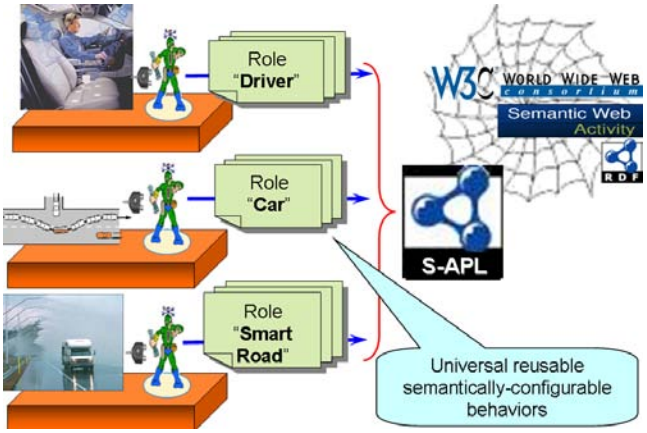



Figure 3. Reusable behaviors for UbiRoad "actors"

## C. *Coordination of UbiRoad Resources*

The coordination layer is the next important layer of GUN and UbiRoad architecture and it is designed to make every domain resource collaborative, which means on the one hand coordination of autonomous and proactive parts of this resource (which are also smart resources themselves) and on the other hand coordinate own behavior with other resources within an organization (or within a scenario, which involves several individual proactive participants as shown in Figure 4) towards achieving consensus between personal and collaborative goals.

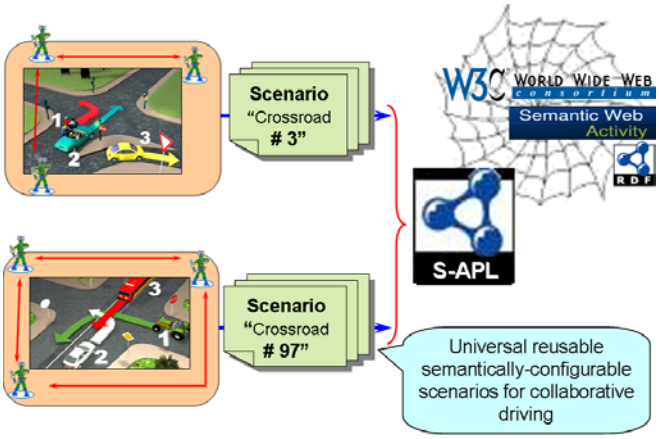



Figure 4. Reusable coordination scenarios in UbiRoad

We designed RpiDF (Resource Process/Integration Description Framework) as S-APL subset and a tool for semantic annotation of policies and metarules for controlling individual behaviors of the resources towards achieving collaborative goals. It extends RDF by making it more suitable for semantic annotation of collaborative behavior of the resources. The extension allows putting explicit constrains on individual rules, plans and utilized atomic behavioral software components, which are intended to implement corroborative goal-driven behaviors (scenarios) of the group of proactive resources (see Figure 4). It should provide ontologies and tools to design, share, reuse and integrate universal semantically-configurable scenarios for required coordination [16].

## V. UTILIZATION OF THE UBIWARE PLATFORM'S FEATURES FOR UBIROAD

UBIWARE ("Smart Semantic Middleware for Ubiquitous Computing") has been developed by Industrial Ontologies Group (http://www.cs.jyu.fi/ai/OntoGroup) according to GUN vision. UBIWARE can be considered as a new software technology and a tool to support design and installation, autonomic operation and interoperability among complex, heterogeneous, open, dynamic and self-configurable distributed industrial systems, and to provide following services for system components: coordination, collaboration, interoperability, data and process integration. The UBIWARE project was a major step in a longer path that aims to build GUN (Global Understanding Environment). That is, a platform or middleware that supports flexible integration of all kinds of resources that have not been a priori designed to be interoperable into new processes that have not been specified when designing the platform. The basic approach in development has been that of agile development – creation of a succession of prototypes with improving functionalities on every release combined with concrete use cases with companies.

The version UBIWARE 3.0 of the platform (Spring-Summer, 2010) appear to be a tool for creating and executing configurable distributed systems based on generalized and reusable business scenarios, which heterogeneous components (actors) are not predefined but can be selected, replaced and configured in runtime. Possible UbiRoad-related scenario on top of UBIWARE 3.0 platform is shown in Figure 5.

Extended version UBIWARE 3.1 (Summer-Fall 2010) is based on Cloud Computing architecture and provides both ontology-driven component- and scenario-based application design and configuration environment for the end-users and also platform-as-a-service to enable continuous run of the applications.

Several innovative features, technologies and components of the UBIWARE platform are making it as an excellent tool to enable the UbiRoad vision and appropriate software implementation. Therefore we will provide more details about it within the following text.

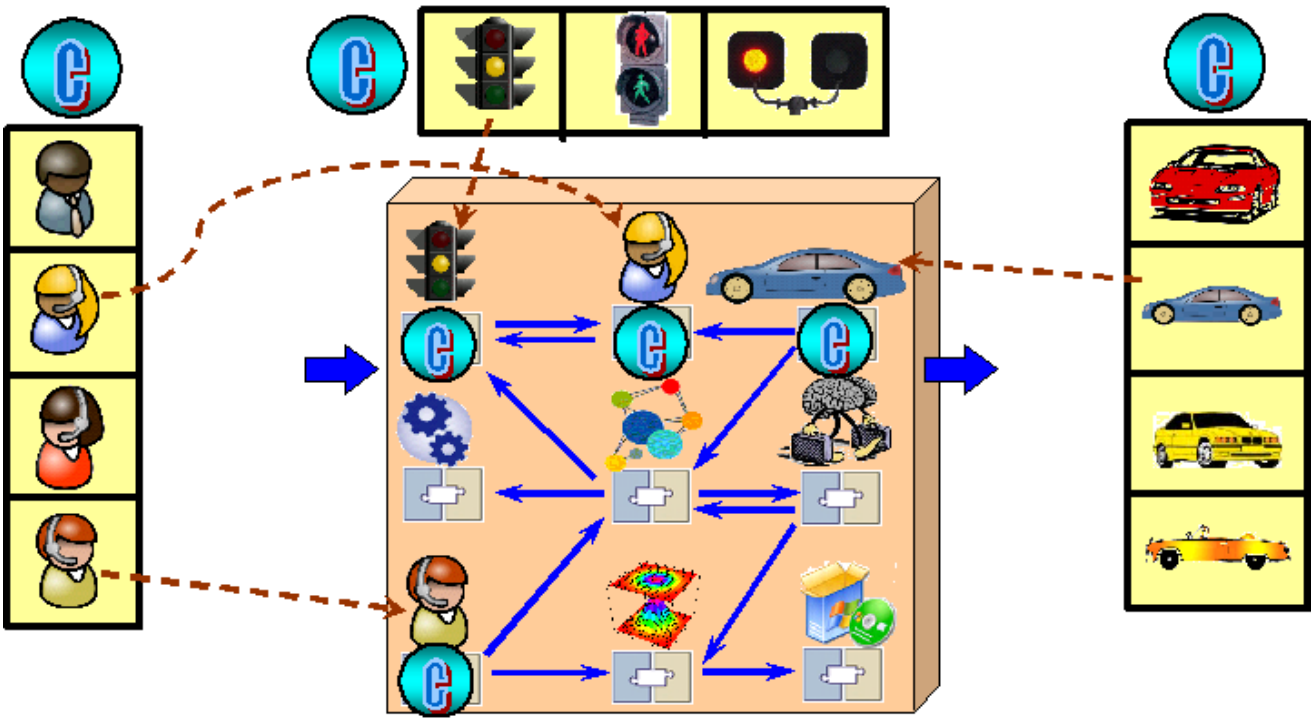


Figure 5. Abstract UbiRoad scenario implemented as a (self)configurable system on top of UBIWARE 3.0 platform

### A. *OntoNuts – Proactive Semantic Adapters*

OntoNuts [22] have been proposed as an ontology-based instrument to enhance complex distributed systems by automated discovery and linking external sources of heterogeneous and dynamic data and capabilities during system runtime.

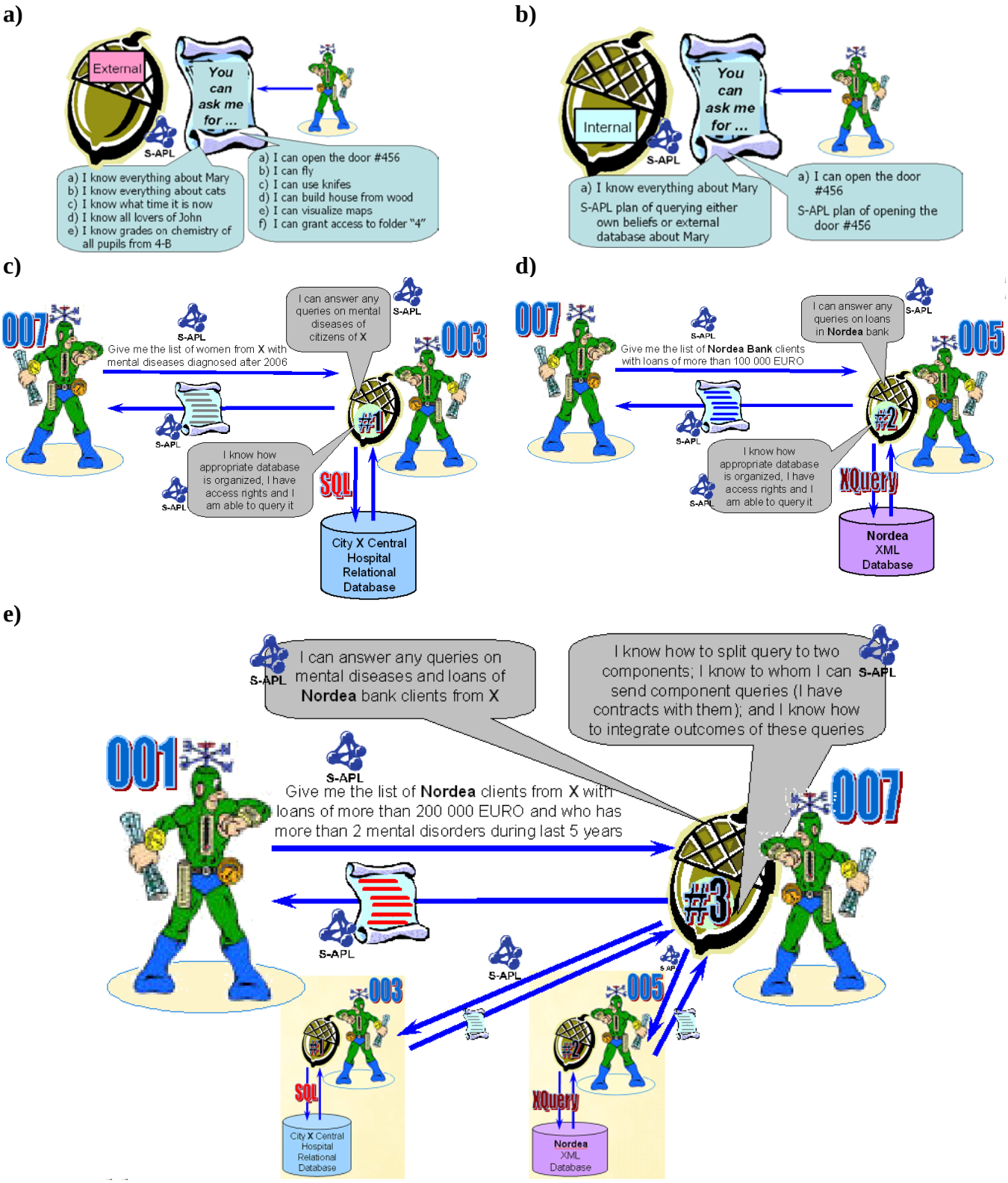



Figure 6. External (a) and internal (b) view on an OntoNut and examples of OntoNuts usage (atomic (c) and (d) and composite (e) OntoNuts)

An OntoNut can be seen from the two points of view and therefore it has two basic components: external and internal. External view to an OntoNut sees it as a tool for proactive advertising of capabilities of some external information (Figure 6 (a)) or service (Figure 6 (b)) provider. By such way each agent is able to actively advertise its capabilities (especially ones related to utilization of external services and databases) to other agents at the platform. Such advertisements generally include semantically annotated capability profile (presented in S-APL). Internal view to an OntoNut contains semantic description (S-APL) of the capability utilization plan for the agent with all needed information on how to access, invoke and monitor process of querying, executing and integrating of some external information sources or services.

Figure 6 (c) (also d) shows example of the OntoNut, which has been designed to wrap the capability of some external database querying. Due to such OntoNut the agent who has direct access to the database and knows how to make appropriate queries to get information from it, now will be also able to advertise such capability to other agents and provide appropriate service for them.

Interesting case is shown in Figure 6 (e) where complex OntoNut wraps two other OntoNuts and therefore is able to perform complex distributed query to two remote and heterogeneous (!) databases.

OntoNuts can be either preprogrammed by system designers or automatically created by agents themselves. Possible general rule for an agent of automatic OntoNut appearance can be presented like below:

**IF** I have the plan how to perform certain complex or simple action or the plan how to answer complex or simple query…

**AND** {time-to-time execution of the plan is part of my duty according to my role (commitment) **OR** I am often asked by others to execute action or query according to this plan}…

**THEN** I will create ONTONUT which will make my competence on this plan explicit and visible to others

OntoNuts approach looks similar to the Semantic Web Services [18] however provides much more potential due to agent-driven proactivity of services (or service components) and data sources. Added value of proactivity for similar purposes has been described in [19].

### *B. 4i (For-Eye) – Visualization-as-a-Service*

4i (For-Eye) [23-24] is a smart ontology-based visualization technology able to automatically discover and utilize external visualization service providers and dynamically create and visualize mashups from external data sources in a context-driven way.

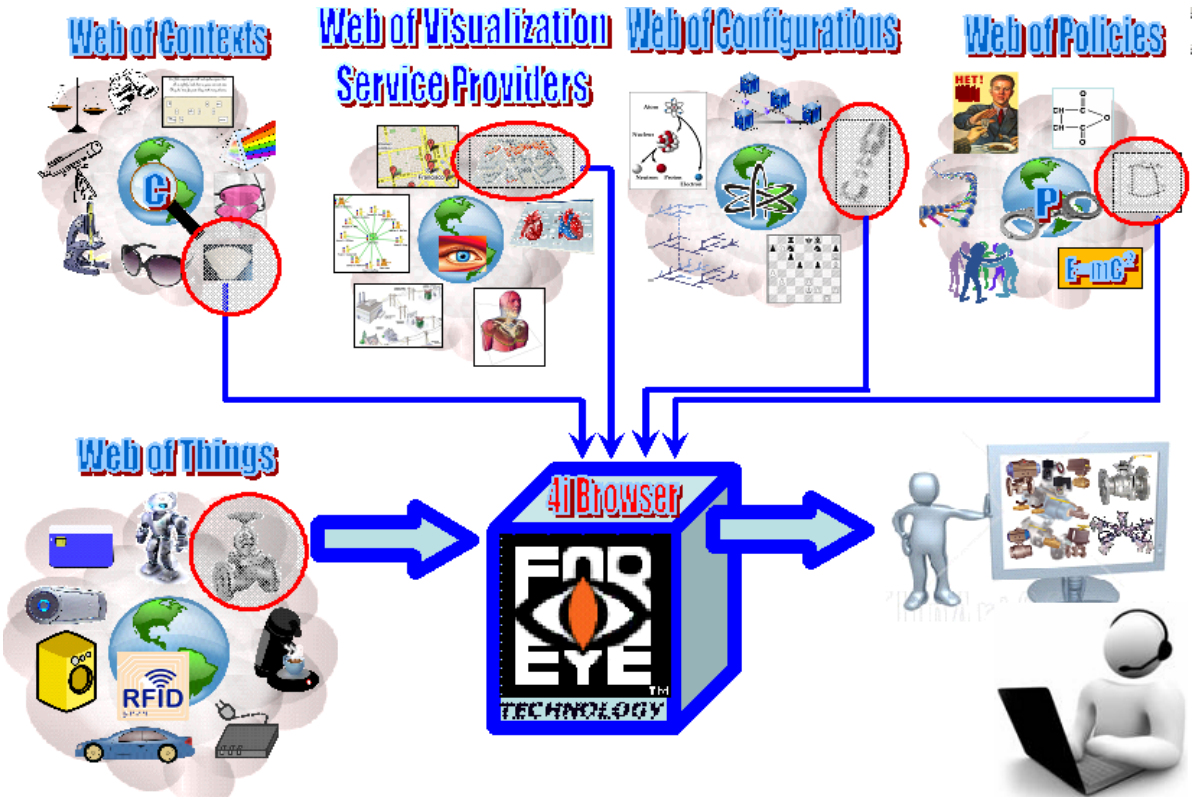



Figure 7. For-Eye-Browser illustrated: Visualization-as-a-Service applied to ubiquitous objects

While OntoNuts are used for advertising, remote access and utilization of external capabilities (software or ubiquitous), 4i technology is applied to enable "Human-as-a-Service" by providing both: interface from UBIWARE infrastructure to a human and vice versa. The key of 4i approach is that instead of designing human interfaces (which is quite labor-consuming task) for each needed combination of data sources, presentation, context, etc, it is proposed to utilize (reuse and integrate when appropriate) external interfacing software components acting as visualization service providers. Therefore slogan of 4i technology is "Visualization-as-a-Service".

The generic vision of future tool (4i Web Browser), restricted version of which is currently part of UBIWARE platform, is shown in Figure 7. Given URI of some resource (document, device, human, etc.) and the task is to visualize it. The selection of properties (to be shown) of the resource as well as the neighborhood of it depend on the context of the concrete viewer. Browser should be aware about the context (either by explicit provision of it from the user or by referencing to URI of some annotated reusable context published in the Web). Also the configuration of the resource may be explicitly provided similarly to the context; and the policy applied to current visualization (which components or properties of the resource can be shown to such a user with particular access rights and which not). Finally appropriate visualization service will be discovered and utilized, which is able to show such kind of resources and their neighborhood on the screen (Figure 7).

### C. *Smart Comments – User-Driven Configuration Tool*

Smart Comments – is smart ontology-based technology for end-user-driven control and configuration management of the application in runtime based on smart mapping of appropriate tags from natural language comments provided by a SW engineer and the source code.

Via Smart Comments an end user (not a programmer) will be able to modify the business logic of an application. Figure 8 illustrates such possibility. If S-APL programmer at the design phase leaves a Smart Comment (natural language description synchronized with the code through several configurable variables) attached to some S-APL construct (e.g. S-APL rule as shown at the picture), then the interface for this rule modification can be automatically generated and called during runtime by the end-user. S-APL modifications can be applied immediately to the constantly running system without stopping it.

Therefore Smart Comments are serving both for documenting S-APL code (supported by UBIWARE engine) and for automatic generation of end-user interface for system reconfiguration whenever needed on the fly.

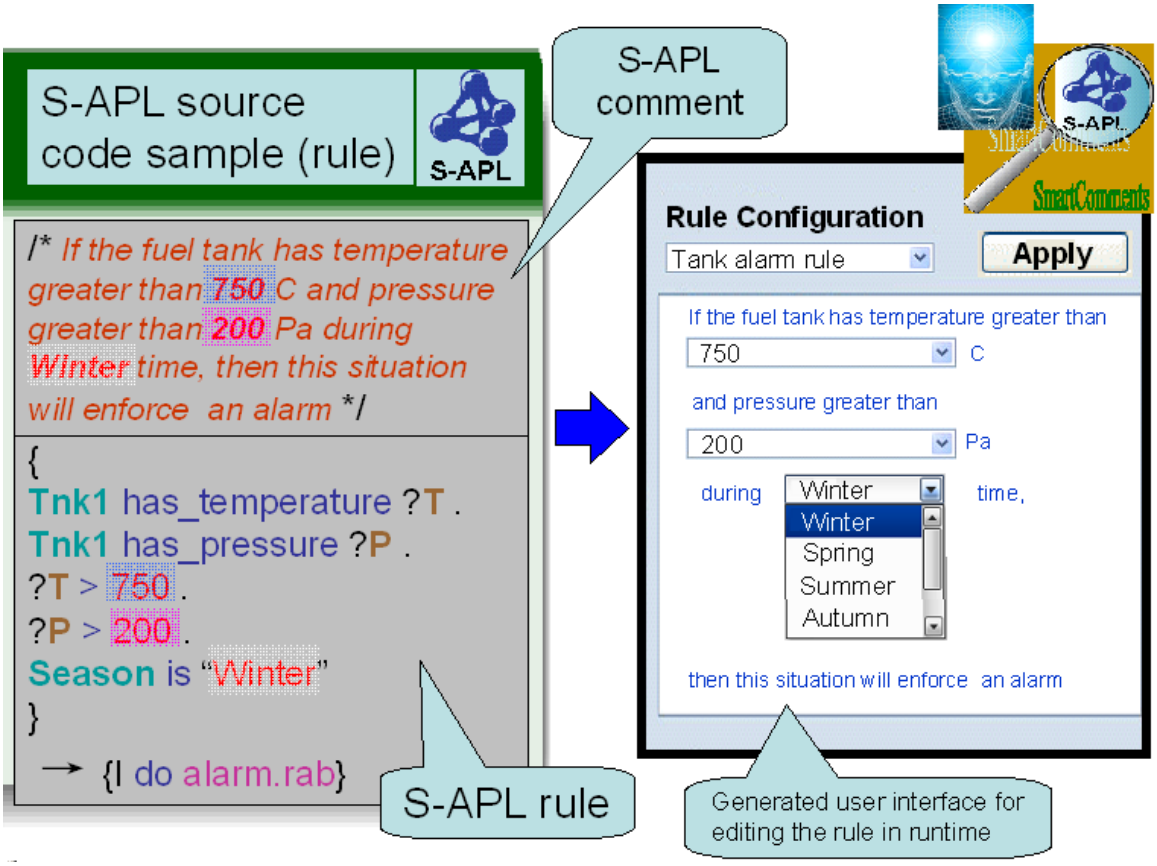



Figure 8. Example of a Smart Comment and appropriate reconfiguration user interface

### D. Semantic Blogging – Collecting Annotated History

Blogging or collecting and publishing own history (usually by humans) is very popular feature of e.g. various Web 2.0 applications. Taking into account that UBIWARE and especially UbiRoad applications are supporting complex distributed business processes, which involve various components (humans, vehicles, devices, services, infrastructure, etc.), we assume that blogging may serve not only to humans but to other ubiquitous or software entities. Own history (especially semantically annotated one) collected by each system component can be later processed by various intelligent tools and useful patterns can be discovered and reused.

In Figure 9, the semantic blogging is shown for the maintenance lifecycle management of a vehicle. Integrated information from sensors, fault detectors, diagnostic software or humans, maintenance workers, etc., collected in timely fashion and automatically annotated provides valuable source of data for predictive maintenance of that particular vehicle and possibly of the same type of vehicles. In UBIWARE, semantic blogging is agent-driven and all history data is represented in S-APL (i.e. can be processed and exchanged by agents).

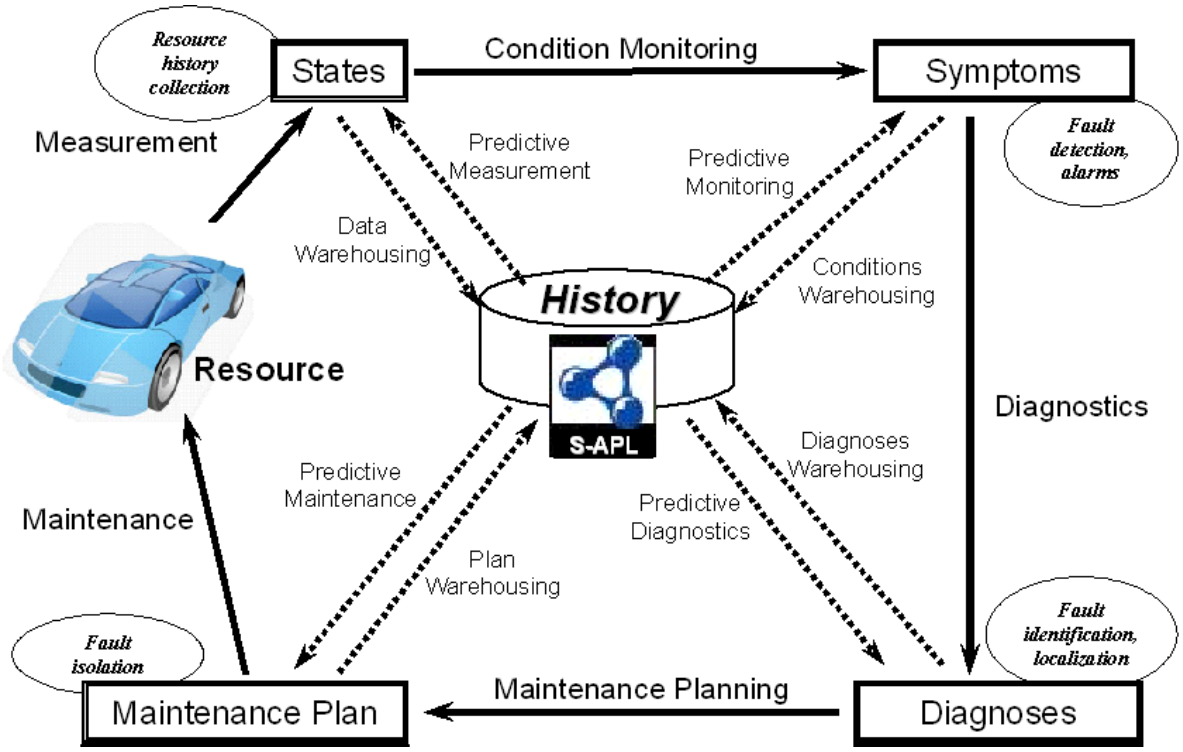



Figure 9. Lifecycle of a semantic blog related to vehicle maintenance domain

## VI. UBIWARE: INNOVATIVE SOFTWARE ARCHITECTURE

The UBIWARE Platform is a development framework for creating complex self-managed multi-agent systems in various domains (not only for the UbiRoad). It is built on the top of the Java Agent Development Framework (JADE), which is a Java implementation of IEEE FIPA specifications. JADE provides communication infrastructure, agent lifecycle management, agent directory-based discovery and other standard services. In UBIWARE, a multi-agent system is seen, first of all, as a middleware providing interoperability of heterogeneous resources and making them proactive and in a way smart. The central to the UBIWARE Platform is the architecture of a UBIWARE agent depicted in Figure 10 together with four main innovations behind it (approach, engine, language, OntoNuts). It can be seen as consisting of three layers: the Behavior Engine implemented in Java, a declarative middle-layer (Behavior Models corresponding to different roles the agent plays), and a set of sensors and actuators which are again Java components. The latter we refer to as Reusable Atomic Behaviors (RABs). We do not restrict RABs to be only sensors or actuators, i.e. components concerned with the agent's environment. A RAB can also be a reasoner (data-processor) if some of the logic needed is impossible or is not efficient to realize with S-APL, or if one wants to enable an agent to do some other kind of reasoning beyond the rule-based one. Current version of UBIWARE platform (see updates in [17]) contains a set of RABs and the libraries that simplify UBIWARE application development.

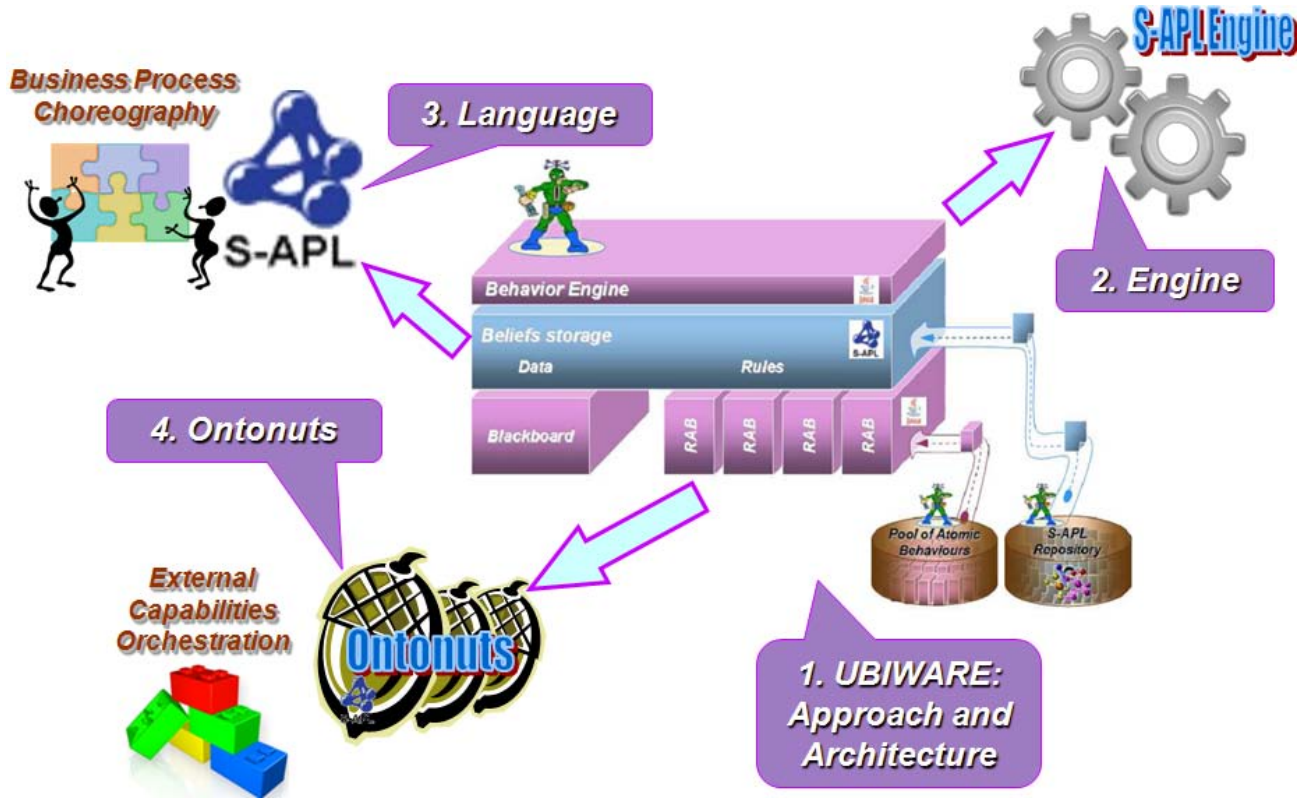



Figure 10. UBIWARE agent architecture with four main innovations

The middle layer is the beliefs storage in Semantic Agent Programming Language (S-APL), which is a Resource Description Framework (RDF) - based language integrating features of several kinds of tools: agent programming languages (like AgentSpeak and AFAPL), semantic reasoners (like CWM), querying languages (like SPARQL), business process description languages (like BPEL) and agent communication content languages (like FIPA SL). What differentiates S-APL from traditional APLs is that S-APL is RDF-based. This provides the advantages of the semantic data model and reasoning. An additional advantage is that in S-APL the difference between the data and the program code is only logical but not any principal. Data and code use the same storage, not two separate ones. This also means that: a rule upon its execution can add or remove another rule, the existence or absence of a rule can be used as a premise of another rule, and so on. None of these is normally possible in traditional APLs treating rules as special data structures principally different from normal beliefs which are n-ary predicates. S-APL is very symmetric with respect to this – anything that can be done to a simple statement can also be done to any belief structure of any complexity.

Together with the main UBIWARE engine, an OntoNuts technology and OntoNuts engine have been implemented. OntoNuts technology tackles the problem of distributed querying in UBIWARE-based multi-agent systems. Due to it and based on the Open World Assumption (alternatively to e.g. BPEL with the Closed World Assumption), the OntoNuts engine is able to discover and automatically utilize (in a runtime) new services (capabilities) that have just appeared in semantic service registry or/and to easily connect external data sources and run distributed queries over them. The backward chaining algorithm was implemented to meet the platform-specific features and the language. The algorithm implementation is used in the planning of distributed queries. To support database connectivity, a special type of OntoNut called DoNut has been implemented that provides additional functionality to the user when dealing with the relational data sources. Special attention was paid to the mapping and transformation (adaptation) of the external sources. It is known that the Service Oriented Architecture (SOA) is an approach of integrating available enterprise applications in a flexible and loosely coupled manner to enable more sophisticated, complex and distributed applications. SOA is built on the notion of services (external capabilities, which are realizations of self-contained business functions). SOA is based on choreography and orchestration of services. Choreography is concerned with describing the external visible behavior of services, as a set of message exchanges, from the functionality consumer point of view. Orchestration deals

with describing how a number of services, two or more, cooperate and communicate with the aim of achieving a common goal. S-APL is a language capable to describe both choreography and orchestration (through OntoNuts) of external capabilities (data or functional services) and internal atomic capabilities (Reusable Atomic Behaviors) needed for designing and executing a complex business process. UBIWARE platform nowadays is such a middleware solution that combines the features of the application server, the semantic web platform and the agent-driven platform, where agent-driven semantic applications can serve end-customers with the high quality web-based GUIs, enhanced user-friendliness and responsiveness. The platform has become an application-independent runtime environment, where special infrastructure agents take care of the platform itself, not of the applications being run on it. At the same time, the personal user agents have been introduced, thus making the platform user-oriented infrastructure for creation of various kinds of applications. Those applications have a freedom to use a web front-end, on-the-platform user management and other infrastructure or define their own platform components depending on the needs of the application. Latest architecture of the platform follows cloud computing paradigm combined with agent architecture, in which two groups of agents were identified. The first group includes the agents which are application-specific, whereas the second group gathers infrastructure agents providing services to those application-specific ones.

## VII. Ontology and Semantic Integration for Traffic Management

To enable interoperable scenarios on top of UBIWARE platform (written in S-APL) and seamless integration of external capabilities there is a need for shared domain ontology. To run UbiRoad application on UBIWARE, the Traffic & Mobility Ontology is being developed as collaborative effort of Industrial Ontology Group, VTT (Technical Research Center of Finland) and Cooperative Traffic ICT SHOK Consortium. Due to complexity of the domain and heterogeneity of components, standards and actors there, such effort is a quite challenging task which will include:

- Vehicles Ontology
- Drivers Ontology
- Infrastructure Ontology
- Logistics Ontology
- Organizations/Products/Services Ontology
- Behavioral Ontology
- Monitoring/Diagnostics/Control/Maintenance Ontology
- Cooperative Scenarios Ontology
- Policy Ontology (security, privacy, safety, economic, skills/demands, environmental, operational, institutional, personal, cultural, etc.)

Ontology is especially important for the OntoNuts technology performance because it allows externalizing not only data sources, services and other capabilities, but also remote and heterogeneous systems as whole.

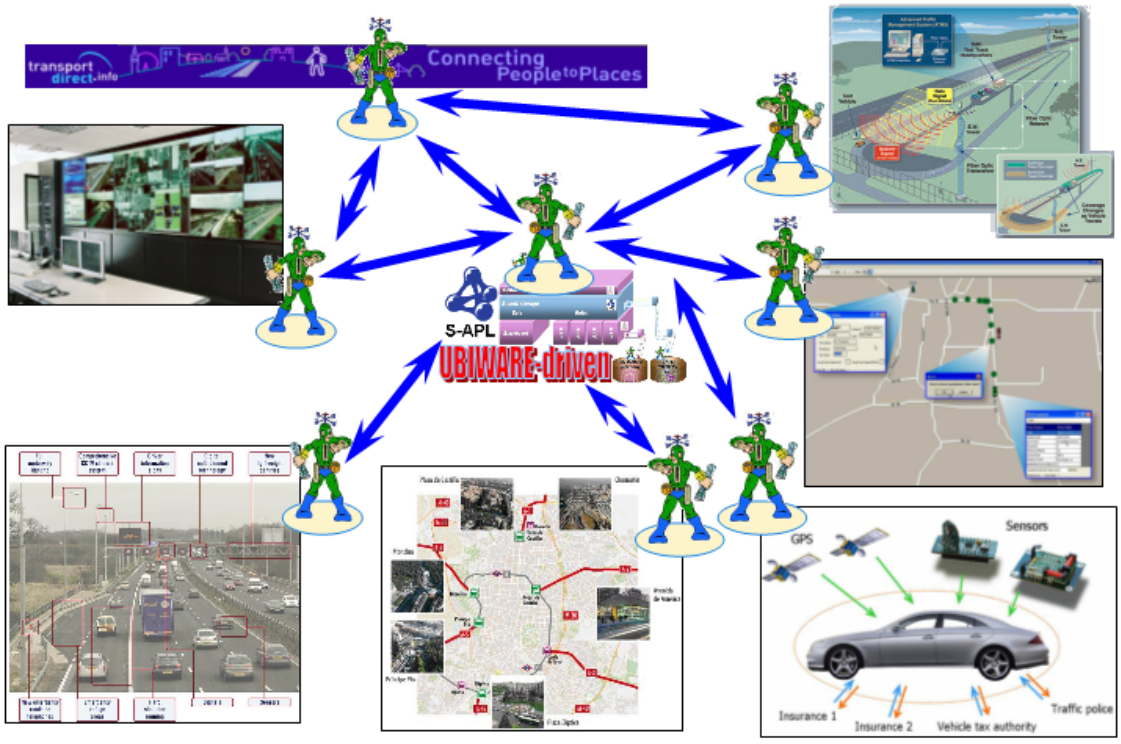



Figure 11. Agent-mediated heterogeneous traffic management systems integration

In Figure 11 it is shown that services provided by remote traffic management systems (heterogeneous, distributed, "self-interested") can be automatically advertised, mediated, utilized and integrated via agent-driven OntoNuts. UBIWARE platform in this case will act as kind of smart semantic "glue" linking and integrating remote systems as services according to user-driven scenarios.

## VIII. RELATED WORK

The topics related to cooperative traffic, traffic management systems, smart traffic environments, driver assistance, etc., are quite popular nowadays. It would be quite challenging to observe all various approaches and solutions in these fields. However not all of them recognize the needs and benefits, which semantic and agent technologies may bring to provide scalability, flexibility and interoperability for available solutions, systems and products. Lets observe few additional samples of research efforts, which are also basing their solutions on semantic or/and agent technology.

The Intelligent Systems Group's (ISG) (http://www.ee.oulu.fi/research/isg) is trying to develop enhanced adaptivity and context-awareness for smart environments. The research specifically focuses on the creation of dynamic models that enable monitoring, diagnostics, prediction and control of target systems (living and artificial) or operating environments making the environment adapt to the users, instead of making the users adapt to an environment. In [25] ISG present their work towards achieving context-awareness in mobile devices by combining Semantic Web technology with sensory data. They show that some context data pertaining to the user, such as location, time, and physical surroundings, is vital for the realization of intelligent maps able to reason on the sensory data with the help of appropriate ontologies and to utilize its inference output to achieve context-awareness.

Researchers from AI & CS Lab, University of Porto, in [26] explored potential benefits of concepts such as visual interactive modeling and simulation to implement a cooperative network editor embedded in a collaborative environment for transport analysis. They argue that traditional approaches lack adequate means to foster integrated analyses of transport systems either because they are strict in terms of purpose or because they do not allow multiple users to dynamically interact on the same description of a model. They show that the use of semantic approach and a common geographical data model of the application domain enable different experts to interact seamlessly in a collaborative environment.

Efforts of NEARCTIS (FP7-th Network of Excellence for Advanced Road Cooperative traffic management in the Information Society, http://www.nearctis.org/) focus on cooperative systems for road traffic optimization, and it covers a wider scope as it appears that cooperative systems have to be integrated into the whole traffic management systems. One of challenges they recognized is to provide means for sharing resources (data, experimental tools, bibliographical databases), organizing the spreading of the knowledge and research results, for which one cannot avoid utilization of Semantic Technologies.

Deducing spatial knowledge for car driver assistance is of special importance for emerging Advanced Driver Assistance Systems. Such systems cannot rely only on in-car sensors infrastructure, but also require thorough environmental tracking. In [27] an approach is presented of a distributed ad-hoc infrastructure that collects and disseminates tracking data of environmental objects and allows ontology-based reasoning. It is shown that such a system can facilitate driver assistance based on spatial knowledge.

In general, driver assistance system demands a common domain understanding for scene representation to enable information exchange between a vehicle and a driver. In [28] an ontology modeling approach is presented for assisting drivers through safety alerts during time critical situations. Designed Intelligent Driver Assistance System (I-DAS) manages appropriate alert parameter representation in XML format while the recognition and interpretation of a critical situation is done using ontology. Authors argue about the feasibility of combining the advantages of ontology with the reasoning power of logic-based languages.

Researchers from Advanced Highway Maintenance and Construction Technology Research Center, University of California-Davis, are investigating issues related to transportation asset management and related infrastructure maintenance. In [29] they present their asset management solution based on combination of semantic models of mobile and stationary transportation assets with the visualization capabilities of Google Earth. Semantic models can represent complex relationships between diverse asset classes and Google Earth is used for visualization because of its accessibility to a wide range of users and ability to combine different types of data. The model defines stationary and mobile assets, and real-time traffic sensors. Results show that the developed semantic models facilitate integration of appropriate software and hardware systems.

Vehicular traffic in modern cities makes our mobility there quite time and resource consuming. Therefore we expect from future traffic management systems significant savings of fuel and time if traffic control mechanism could be effectively discovered.

In [30] the problem of real time traffic data availability and processing is handled by utilizing ubiquitous database and intelligent agents for traffic data management. Ubiquitous database provides automatic everywhere access to the data and so the called unique routing agent is used to handle the distribution of the database, route discovery and maintenance. The method has been simulated for the measurement of traffic related parameters (traffic load, occupancy and trip time).

## IX. CONCLUSIONS

In this paper, we approach the traffic-collaboration-support problem from the semantic viewpoint. In other words, the semantic technologies have a two-fold value in UbiRoad. First, they are the basis for the discovery of heterogeneous resources and data integration across multiple domains (a well-known advantage). Second, they are used for behavioral control and coordination of the agents representing those resources (a novel use). Therefore, semantic technologies are used both for descriptive specification of the services delivered by the resources and for prescriptive specification of the expected behavior of the resources as well as the integrated system (i.e., declarative semantic programming). While the standard semantic technology is capable of effective description of static resources only, the UbiRoad is a tool for semantic management of content relevant to dynamic, proactive, and cooperative resources. The agent technology is extended by developing tools for semantic declarative programming of the agents, for massive reuse of once generated or designed plans and scenarios, for agent coordination support based on explicit awareness of each other's actions and plans, and for enabling flexible re-configurable architectures for agents and their platforms applied for cooperative traffic domain. However, taking into account that the efficiency of semantic technologies and available tools for real-time applications as well as agent technologies (e.g., agent negotiation within fast developing situations) is still questionable, a smart way to combine semantic and agent approaches with efficient online data processing and automation tools would be reasonable.

This paper is an extended version of conference paper [1] accepted for journal publication.

## REFERENCES

[1] Terziyan, V., Kaykova, O., and Zhovtobryukh, D., UbiRoad: Semantic Middleware for Context-Aware Smart Road Environments, In: Proceedings of the Fifth International Conference on Internet and Web Applications and Services (ICIW-2010), May 9-15, 2010, Barcelona, Spain, IEEE CS Press, pp. 295-302.

[2] Hype Cycle for Automotive Integration and Communication Technologies, 2006, Gartner.

[3] Chen-Ritzo, C-H, C. Harrison, J. Paraszczak, and F. Parr. "Instrumenting the Planet." IBM Journal of Research & Development. Vol. 53. No. 3. Paper 1. 2009.

[4] Cooperative Traffic ICT: Strategic Research Agenda, May 2009, ICT SHOK, Finland, Available in: http://www.cooperativetraffic.fi/ images / 1 /16/SRA_CT_v2.pdf .

[5] Buckley, J., From RFID to the Internet of Things: Pervasive Networked Systems, Final Report on the Conference organized by DG Information Society and Media, Networks and Communication Technologies, Brussels, 2006. Available in: http://www.rfidconsultation.eu/docs/ficheiros/WS_1_Final_report_27_Mar.pdf .

[6] Kaykova, O., Khriyenko, O., Kovtun, D., Naumenko, A., Terziyan, V., and Zharko, A., General Adaption Framework: Enabling Interoperability for Industrial Web Resources, International Journal on Semantic Web and Information Systems, 1(3), 2005, pp. 31-63.

[7] Terziyan, V., Semantic Web Services for Smart Devices Based on Mobile Agents, International Journal of Intelligent Information Technologies, Idea Group, 1(2), 2005, pp. 43-55.

[8] Terziyan, V. SmartResource – Proactive Self-Maintained Resources in Semantic Web: Lessons learned, In: International Journal of Smart Home, Special Issue on Future Generation Smart Space, Vol.2, No. 2, April 2008, pp. 33-57.

[9] Berners-Lee, T., Hendler, J., and Lassila, O., The Semantic Web. Scientific American, 284(5), 2001, pp. 34–43.

[10] Katasonov, A., Kaykova, O., Khriyenko, O., Nikitin, S., and Terziyan, V., Smart Semantic Middleware for the Internet of Things, In: Proceedings of the 5-th International Conference on Informatics in Control, Automation and Robotics, 11-15 May, 2008, Funchal, Madeira, Portugal, pp. 169-178.

[11] Terziyan, V., Semantic Web Services for Smart Devices in a "Global Understanding Environment", In: R. Meersman and Z. Tari (eds.), On the Move to Meaningful Internet Systems 2003: OTM 2003 Workshops, Springer-Verlag, LNCS, 2889, pp. 279-291.

[12] Terziyan, V. and Katasonov, A., Global Understanding Environment: Applying Semantic and Agent Technologies to Industrial Automation, In: M. Lytras and P. Ordonez De Pablos (eds.), Emerging Topics and Technologies in Information Systems, IGI Global, 2009, pp. 55-87 (Chapter III).

[13] Katasonov, A. and Terziyan, V., Implementing Agent-Based Middleware for the Semantic Web, In: Proceedings of the International Workshop on Middleware for the Semantic Web in conjunction with the Second IEEE International Conference on Semantic Computing (ICSC-2008), August 4-7, 2008, Santa Clara, CA, USA, IEEE CS Press.

[14] Kaykova, O., Khriyenko, O., Naumenko, A., Terziyan, V., and Zharko, A., RSCDF: A Dynamic and Context-Sensitive Metadata Description Framework for Industrial Resources, Eastern-European Journal of Enterprise Technologies, 3(2), 2005, pp. 55-78.

[15] Kaykova, O., Khriyenko, O., Terziyan, V., and Zharko, A., RGBDF: Resource Goal and Behaviour Description Framework, In: M. Bramer and V. Terziyan (Eds.): Industrial Applications of Semantic Web, Proceedings of the 1-st International IFIP/WG12.5 Working Conference IASW-2005, August 25-27, 2005, Jyvaskyla, Finland, Springer, IFIP, pp. 83-99.

[16] Katasonov, A. and Terziyan, V., Semantic Approach to Dynamic Coordination in Autonomous Systems, In: R. Calinescu et al. (Eds.), Proceedings of the Fifth International Conference on Autonomic and Autonomous Systems (ICAS 2009), April 21-25, 2009, Valencia, Spain, IEEE CS Press, pp. 321-329.

[17] Web Pages of the Industrial Ontologies Group: Available in: http://www.cs.jyu.fi/ai/OntoGroup/index.html .

[18] Arroyo, S., Lara, R., Gomez, J., Berka, D., Ding, Y., and Fensel, D., Semantic Aspects of Web Services: Practical Handbook of Internet Computing, Chapman & Hall and CRC Press, USA, 2004, 31.31-31.17.

[19] Ermolayev, V., Keberle, N., Plaksin, S., Kononenko, O., and Terziyan, V., Towards a Framework for Agent-Enabled Semantic Web Service Composition, International Journal of Web Service Research, Idea Group, 1(3) , 2004, pp. 63-87.

[20] Bell, D., Heravi, B., and Lycett, M., Sensory Semantic User Interfaces (SenSUI), In: Proceedings of the International Workshop on Semantic Sensor Networks, 2009, pp. 96-109.

[21] Luther, M., Liebig, T., Bohm, S., and Noppens, O., Who the Heck is the Father of Bob?, In: Proceedings of ESWC-2009, Springer, LNCS 5554, pp. 66-80.

[22] Nikitin, S., Katasonov, A., and Terziyan, V., Ontonuts: Reusable Semantic Components for Multi-Agent Systems, In: R. Calinescu et al. (Eds.), Proceedings of the Fifth International Conference on Autonomic and Autonomous Systems (ICAS 2009), April 21-25, 2009, Valencia, Spain, IEEE CS Press, pp. 200-207.

[23] Khriyenko, O., and Terziyan, V., A Framework for Context-Sensitive Metadata Description, In: International Journal of Metadata, Semantics and Ontologies, Inderscience Publishers, 2006, Vol. 1, No. 2, pp. 154-164.

[24] Khriyenko, O., "4i (FOR EYE) Multimedia: Intelligent Semantically Enhanced and Context-Aware Multimedia Browsing", In: Proceedings of the International Conference on Signal Processing and

Multimedia Applications (SIGMAP-2007), Barcelona, Spain, 28-31 July, 2007, pp. 233-240.

[25] Su, X., Riekki, J., and Tarkoma, S., An Approach to Achieve Context-Aware Maps: Combining Semantic Web Technology with Sensor Data, In: Proceedings of the 5th International Conference on Intelligent Environments (IE09), Barcelona, Spain, 2009, pp. 193-203.

[26] Pereira, J., Rossetti, R., and Oliveira, E., Towards a Cooperative Traffic Network Editor, In: J. Luo (ed.), Proceedings of the 6th International Conference on Cooperative Design, Visualization, and Engineering (CDVE'09), LNCS, Vol. 5738, Springer, 2009, Luxembourg, pp. 236-239.

[27] Tonnis, M., Klinker, G., and Fischer, J., Ontology-Based Pervasive Spatial Knowledge for Car Driver Assistance, In: Proceedings of the Fifth IEEE International Conference on Pervasive Computing and Communications Workshops (PerComW'07), NY, USA, 2007, pp. 401-406.

[28] Kannan, S., Thangavelu, A., and Kalivaradhan, R., An Intelligent Driver Assistance System (I-DAS) for Vehicle Safety Modeling using Ontology Approach, In: International Journal of UbiComp, Vol.1, No.3, July 2010, pp. 15-29.

[29] Darter, M., Lasky, T., and Ravani, B., Transportation Asset Management and Visualization Using Semantic Models and Google Earth, In: Transportation Research Record: Journal of the Transportation Research Board, Vol. 2024, 2008, pp. 27-34.

[30] Dave, N., and Vaghela, V., Vehicular Traffic Control: A Ubiquitous Computing Approach, In: Communications in Computer and Information Science, Springer, 2009, Vol.40, Part 7, pp. 336-348.

# Design of Cognitively Accessible Web Pages

Till Halbach Røssvoll, Ivar Solheim
Norwegian Computing Center (Norsk Regnesentral), Norway
{*halbach,solheim*}*@nr.no*

***Abstract*—Considering the design of universally designed interfaces for static and dynamic web pages, this work focuses on the group of users with cognitive/intellectual disabilities, while simultaneously accounting for the needs of users with motor and sensory deficits. A number of specific inclusive techniques are applied to the login mechanism of a web service in the course of the redesign of this site. The techniques evolve, i.e. are tested, validated, and refined, over a series of implementation iterations and subsequent evaluation, involving personas and scenario testing, an expert panel, and user testing. The testing shows that the web service's resulting login mechanism is much more universally accessible than today's solution. Generically applyable, universal design principles are derived for a number of intellectual deficits, such as problems with linguistics (text and language), learning and problem solving, orientation, focus and attention span, memory, and visual comprehension.**

***Keywords-Cognitive disabilities; intellectual deficits; impairment; deficiencies; accessibility; e-inclusion; universal design; user experience; web pages.***

## I. Introduction

Accessible web sites and online services is a topic of high concern for industry, public actors, and research likewise. However, people with sensory deficits are typically in focus here, while motor and cognitive impairments often are given less attention. This article concentrates on the latter, sometimes also referred to as intellectual deficiencies, extending the work presented in [1].

There is a strong rationale to address the topic of cognitively accessible web pages. In 2006, roughly 22 million people in the United States were counted to have cognitive disabilities due to various reasons [2], while world-wide estimates for 2008 range as high as 400 million people [3]. These numbers include the intellectual challenges typically encountered by a number of elderly people with various degrees of severity.

The starting point for the project described in this article was a case provided by the Norwegian public services provider Altinn [4], involving a redesign of their site [5]. One of the requirements for the new design and the new functionality was to accommodate for people with cognitive challenges. A detailed survey of all requests to Altinn's help desk had revealed that 33% of the users had problems with the login process [6], and the service provider considered it a strategic goal to reduce the number of help requests by developing a new page design and an improved service architecture.

The paper is organized as follows. After an introduction to relevant cognitive impairments (Section II), a brief review (Section III) of related and previous work summarizes other research, points at gaps to fill, and names this work's contributions towards these goals. Next, the status quo of the current solution is detailed (Section IV) before the development method is explained (Section V). This is followed by the listing of generic design principles with a subsection for each considered cognitive impairment (Section VI). This also includes a discussion of the benefits for other user groups as well (Section VII). After that, the prototype is presented (Section VIII), with a special section on instruction videos (Section IX). Finally, there is a general discussion regarding the consequences of this work's findings (Section X) before the final conclusion is drawn.

## II. Cognitive impairments

Cognitive impairments can be defined as "(the) substantial limitation of one's capability to think, including conceptualizing, planning, sequencing thoughts and action, remembering, interpreting subtle social cues, and manipulating numbers and symbols" [2] and can appear at any age. Causes for these impairments include prenatal fatal influences, injuries, and mental illnesses. Basically, a person with cognitive challenges has over-the-average difficulty succeeding with one or more types of mental tasks.

There are several ways to classify cognitive disabilities. In the context of this research it seems appropriate to distinguish between a clinical-diagnostic and a functional approach. Clinical diagnoses of cognitive disabilities include

- autism,
- Down Syndrome,
- traumatic brain injury (TBI),
- dementia,
- dyslexia,
- dyscalculi, and
- learning difficulties in general [7].

Clinical diagnoses are of course helpful and necessary from a medical perspective, but for the purpose of accessibility, classifying cognitive disabilities by functional disability is more useful. Functional disabilities ignore the medial and behavioral causes of the disability and instead focus on the resulting abilities and challenges [7], [8]. It is also worth mentioning that, naturally, any impairment can have various degrees of severity, ranging from mild variants to extreme cases. This vast range makes the universal design of web pages very challenging, and it is obvious that even the most cautious design cannot cope with all variants of impairments that exist.

By looking at the aforementioned survey, areas of difficulty with the previous solution could be identified. E.g., 27% of the calls to Altinn's help desk were related to finding the right service, and 33% of all users had problems with the login routine [6]. The close inspection of data also allowed to classify user groups according to a particular deficit. For instance, the cognitive requirement needed to find the proper service is the ability to orientate in a website.

The following listing of impairments was identified as relevant to represent the target group of users with cognitive deficits.

- Linguistic (text and language)
- Learning and problem solving
- Orientation
- Focus and attention span
- Memory
- Visual comprehension

One of the most important objectives of the project was to derive a number of principles for good web design concerning these given target groups. This process started with a review of related work.

## III. Previous and related work, and this article's contributions

Works related to the topic of this project include an early version of this paper which has been presented previously with preliminary results [1].

In the design guidelines for web design for impaired users, i.e., without a specific focus on the cognitively disabled, [9] give a large number of detailed design instructions. [10] derive a set of cognitive user characteristics based on neuroscience; however, the work lacks concrete design suggestions. While the [11] provides an exhaustive and detailed listing with concrete design principles, it remains unclear how these are justified with regard to the particular cognitive impairments considered. The same applies to the work by [12], even though this work is not equally detailed. Next, [13] presents a great number of design suggestion with a considerable amount of detail, and [7] lists a set of concrete design guidelines in a tabular format and marks them as "applies to" with regard to four major areas of cognitive challenge. On the downside, both documents lack the justification for why particular guidelines are derived. [14], [15] discuss cognitive accessibility with regard to concrete examples and list some derived practical design suggestions. The listings appear to be far from complete, though. And last but not least, parts of the WCAG 1.0 [16] and 2.0 [17] specifications cover measures for cognitive impairments, but only to a limited degree [18], [11], [14].

In contrast to the aforementioned research, the contribution of the present work is to derive generic design principles/guidelines for people with cognitive deficits from concrete examples, and by means of testing and user studies. Moreover, each guideline is associated and hence classified with regard to a specific impairment. Also, in contrast to the cited works, the focus here includes orientation problems as well, as they seem to be very common (27%) with the given (Altinn) case.

## IV. Current solution

Figure 1 shows a screenshot of the current login solution at the time of writing, which was the starting point for the development of the inclusive design. It illustrates several problematic areas, including those discussed below.

1) The grayed-out area on top shows inaccessible functionality, is irrelevant in the given setting, and does thus not help users to focus on the login process below in a satisfactory manner.
2) The user has to choose the preferred login method out of a list of options on the middle left, such as password, PIN code from mobile phone, and smart-card. There may be too many list items to read for users with reading deficits.
3) The little icon, which is placed after each list item, and which symbolizes the security level of each login method, is likely to confuse non-technical users and those with problem solving challenges. It may further be problematic for individuals with visual comprehension deficits.
4) The main login part on the middle right of the page consists of the fields for user input such as social security number and password, and changes according to the choice of login method on the left. The resulting two-column layout of the page may be too complex to understand for users with focus problems, and it might also be problematic to those with attention span challenges.

Also the screenshot of the "My page" shown in Figure 2, at which the user arrives after the login procedure, shows a page with a number of issues, including the following.

1) The page structure is rather complex and not easily comprehensible. It is not straight forward to understand why a particular piece of content is relevant for the user to reach her goals, and how this content relates to other content on the page. This is likely to confuse particularly users with orientation problems and learning difficulties.
2) There is too much information to process, and there is a lot to read for the user to understand the page structure. This might be problematic especially concerning people with dyslexia.

By way of conclusion, it appears the technical possibilities are in the center of today's solution. Content is grayed out because it is a "cool" design effect, too many login methods are presented because — among other reasons — they are technically possible, the security level is shown, even though it is only of interest to the minority of users, and parts of the page are dynamically altered because it is technically possible to embed all parts of one task in the same page. What is needed is a solution which puts the human and his and her needs into the center.

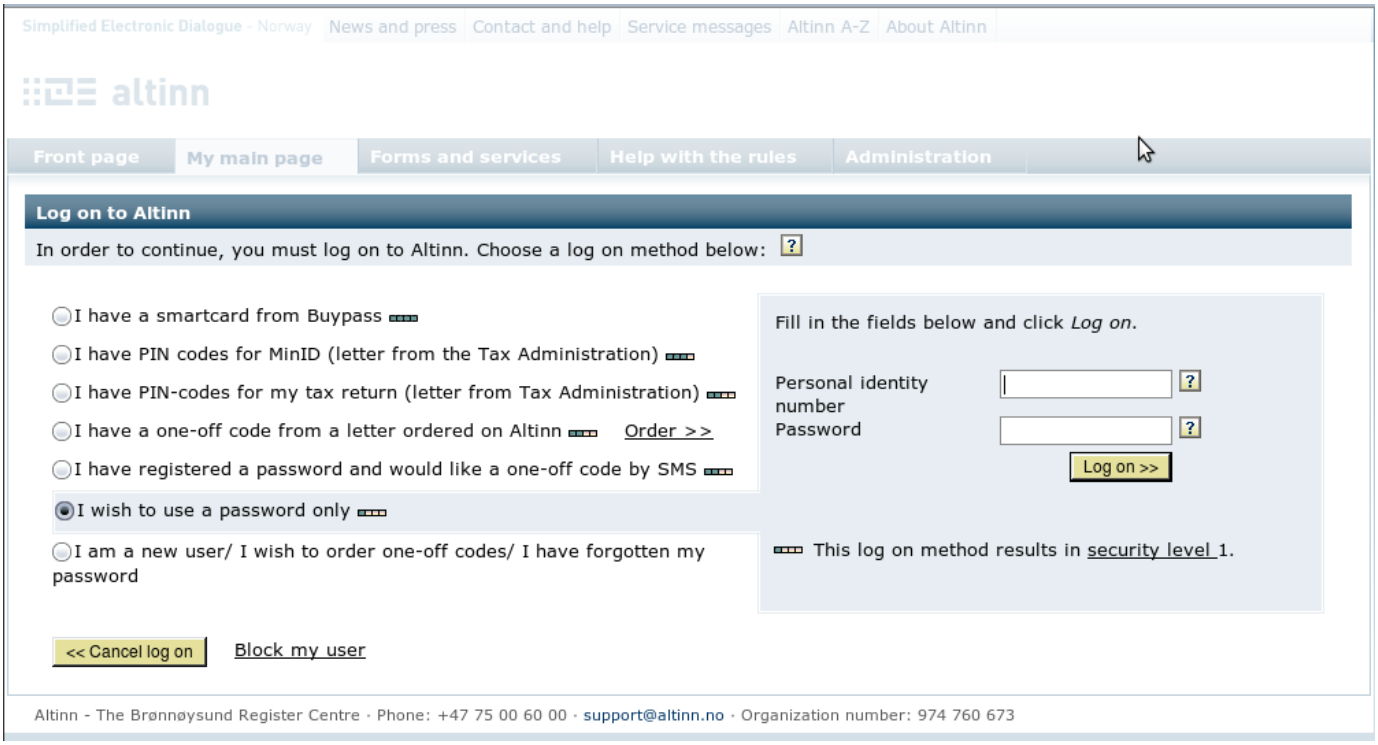



Figure 1. Screenshot of the current login solution

## V. Methodology

Among the objectives of the project were to build a prototype for a new login solution with improved design and functionality as compared to the current solution as specified in Section IV. Another goal was to derive generic guidelines concerning the design of web pages with regard to people with intellectual impairments. It is simultaneously stressed that the developed solution also accounts for the needs of individuals with mobility and sensory deficits, as the user interface meets the requirements of the WCAG 2.0 Recommendation Level AA.

The design guidelines were formulated as hypotheses, implemented in the prototype, and tested for verification. The guidelines are naturally influenced by the results of user testing in other previous projects of our research team. The implementation was refined in several iterative cycles to reflect the feed-back from each evaluation phase.

There were multiple types of evaluation: First, different persona profiles helped to speed up the implementation and testing in the beginning of the development process. The fictive characters, six in total, were given appropriate properties to cover the spectrum of impairments of the target groups, such as “has concentration difficulties”, “poor memorizing abilities”, etc. They were associated with scenarios, allowing simple and cheap cognitive walk-throughs by means of role plays and methods like “thinking aloud”.

Second, when the prototype had reached a certain degree of maturity, the evaluation was conducted by a panel consisting of experts in accessibility, e-inclusion, and universal design. The experts were presented walk-throughs while discussing all aspects of the implementation and were able to provide the latest feed-back from their research areas.

Third, while getting close to the prototype’s completion, a minor user study with eight users representing the cognitive target groups was carried out. The users had various cognitive challenges, such as minor dementia, reading and writing difficulties, focus problems, etc. The number of users was bound by budget limitations. We believe, however, that viewed as a complement to the personas and expert evaluation, the size of the user study is reasonable.

Finally, all design recommendations were collected in an online best-practices tutorial [19]. As the financing institutions of the projects limited the target to the Norwegian market, the tutorial currently comes in Norwegian only, but an English translation is planned in the long-term. Besides topics addressing

- universal design,
- legal matters,
- cognition,
- on system planning, specification, implementation, and evaluation,
- related recommendations, specifications, standards, and standardization organizations,
- useful links and tools,
- glossary, and
- literature

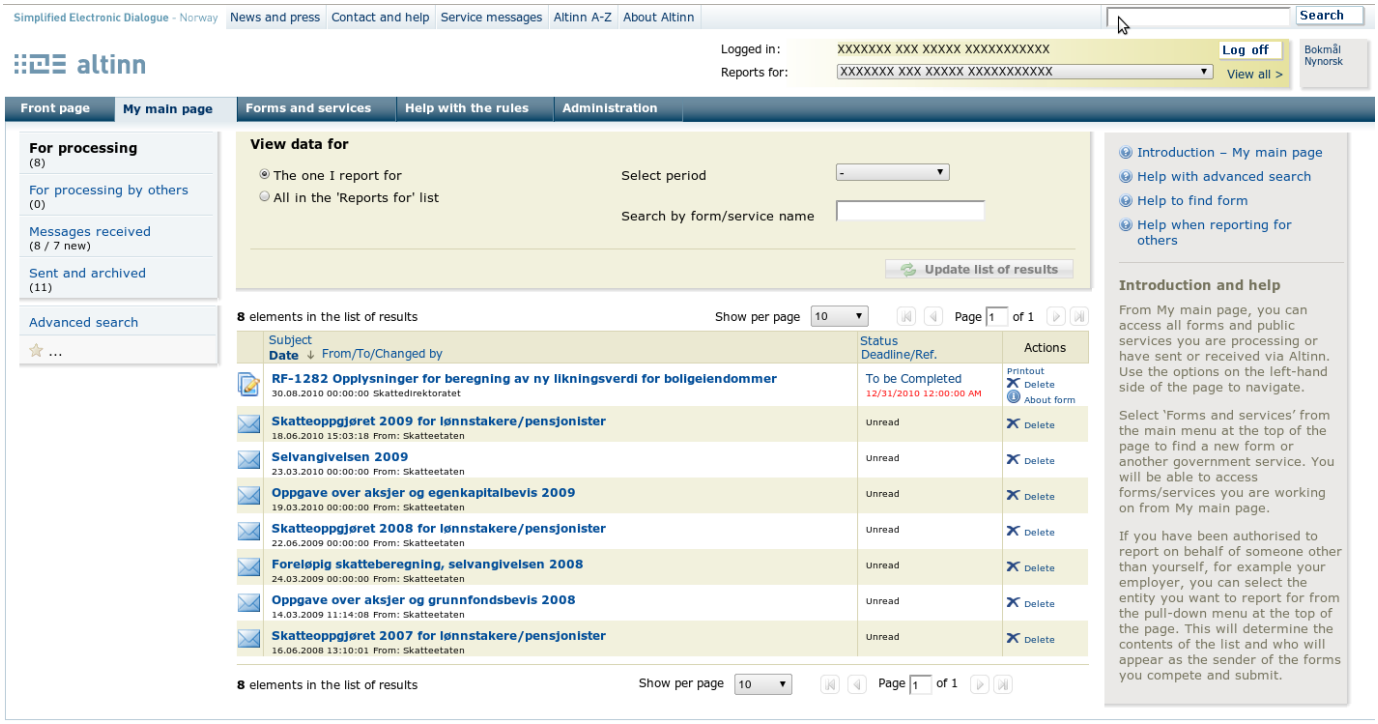



Figure 2. Screenshot of Altinn's "My page" (partly in Norwegian)

the tutorial also lists the design guidelines with practical examples in Hyper-Text Markup Language (HTML), Cascading Style Sheets (CSS) and JavaScript.

## VI. Design Recommendations

The following sections detail the identified design principles or guidelines. They are meant as recommendations and "best practice". Neither of the principles below are listed in any specific order. It is noted that a design measure for a particular functional area may be in conflict with measures from different functional areas, and it therefore happens in some cases that opposite measures for different functional areas have to be balanced against each other.

### A. Text and language

Linguistic problems are difficulties with writing and in particular reading larger amounts of text. A dyslectic may represent this user group. There are no recent statistics, but numbers for Norway from 2005 indicate that approximately 1/3 of the population have moderate to serious reading and writing challenges [20], while other sources talk about 15-20% of the population [21].

The following non-exclusive listing of design principles accommodates linguistic problems.

- Short paragraphs with a reasonable amount of text
- Text in columns with a limited number of characters per line. Concurrent research is yet inconclusive concerning the optimum line length for highest comprehension [22], but we believe that 60–100 characters is a reasonable number.
- Short and concise sentences
- Avoidance of non-literal text, such as allegories, metaphors, slang, and colloquialism
- Avoidance of technical expressions and expert talk
- As few abbreviations and acronyms as possible, and all with proper explanation
- Use of short or non-compound words in languages were single words can be assembled to longer words (such as Norwegian and German)
- Enhancement of semantics by high-quality multimodal content, e.g., symbols/icons, graphics/images, audio, video (depending on the context)
- Textual content structured in short and easily comprehensive logical units like paragraphs and lists, preferably with a heading in advance. Units easily separable from the remaining content
- Choice among several languages for both international and national sites
- Sufficiently long display of subtitles or help text in video to enable slow readers to capture everything. It is of advantage if the user can pause play-back, repeat a timed media sequence, or alter the play-back velocity.

### B. Learning and problem solving & orientation

Some individuals lack the mental flexibility to process information and to apply knowledge in order to solve a given problem. Combined with a low mental endurance, this

typically results in frustration and a user turning away from the task set out. After all, the vast majority of users seeks to get things done with the least possible effort [23] and thus expect things simply to work [24]. Many tasks which have to be solved, such as login, are viewed only as an impediment before the main objective beyond (for instance, filling out an electronic tax statement form) can be accomplished.

Orientation difficulties are part of this class of problems. To cope with these challenges, the design principles listed below should be followed.

- Standard compliant, working solutions, tested thoroughly and on a number of different platforms and user agents (e.g., browsers) to ensure compatibility
- Choice among several alternatives, such as login methods, to solve a task so that a user can pick the one she is most familiar with. However, the number of alternatives should be kept low, depending on the context.
- Multiple modalities for conveyance of content, and possibility to let the user decide upon the preferred modality. An example is to accompany an instruction video by showing a series of key frames from that video with textual description conveying the same message
- Show only content relevant in a given setting to ease orientation
- Common design conventions to make processes predictable, such as hovering effects for responsive user interfaces, known technologies like drop-downs to compact item listings, and for instance a top-bottom left-right ordering of information in terms of relevance for users with a Western background
- Consistent layout and functionality, to give a "learned once, apply everywhere" effect
- Information about the process, like "what" (description), "why" (reason) and "how" (clear instructions), as well as informative error messages and showing potential solutions to problems aim at supporting the process of solving a particular task
- Provision of not only links to help and contact information but also expert systems and demonstrations to make the user's threshold to seek help low enough, and to enable them to help themselves
- Help and demonstration as specific as possible
- Content in logical units which are easily distinguishable from each other
- Information about where in the hierarchy a user currently is, the so-called bread crumb trail
- Responsive user interface, with hints for content and functionality, such as so-called tooltips or other hovering effects on buttons, links, and other page elements, continuation dots for listing extracts, dynamic mouse pointer form depending on the underlying content, prefilled text input fields, etc.
- Classification of large amounts of data with regard to several criteria, with pointers for each classification, to let the user make the preferred mental connection
- Personalized content and functionality in terms of user profiles and sessions, which is essential for functionality like the user's "most used services", "services used last time", and "self-chosen services", as well as state of visit, in terms of data like recognized user name and date and time of last visit
- Ability to search the entire site in an intelligent manner in order to quickly find exactly the information or resource the user has been looking for, for those who prefer to search rather than to navigate
- A 2-step method concerning the user's approach to large quanta of information is often helpful, where a simplified view should be the default option, from which a link to the high-level, i.e., complex, view should be provided
- A personalized "latest news" section, giving service status information, and an update on changes since the last visit and on current important issues
- Avoidance of lengthy scrolling, rather provision of links to additional content
- Links leading to content on the same site should be opened in the same window/tab as the current page. The user should be informed before opening new windows or tabs with content on external sites

### C. Focus and attention span

Attention span refers to the ability to focus on what is important at a particular point in time, and to be able to keep that focus during a longer time period. Similar aspects are one's concentration ability and distractibility.

The design principles identified to accommodate attention span deficits include the following items.

- Only content relevant in a given context, in particular no display of grayed out, irrelevant content
- Use of static page elements, and avoidance of flashing and scrolling elements
- Visual cues to draw the user's attention, such as highlighting the active input field
- Larger processes split up into smaller logical chunks, each of which can be solved with a low attention span
- Consistent layout and page structure in order not to distract the user
- Modifications of the page after it has finished loading not too far away from the center of the page to gain the user's attention
- Avoidance of long durations of timed media, such as an audio clip or a video

### D. Memory

Memory difficulties in general denotes the user's ability to recall what has been learned over time. Any memorizing type can be affected, such as working memory, and short-term and long-term memory. Important design principles accounting for these challenges are listed subsequently.

- Larger processes split up into several logical units/tasks, each of which as brief and simple as possible, according to the Divide-and-Conquer principle
- Reminders concerning the overall context (e.g., "You want to fill out a tax form"), and explanation of the particular context ("Therefore you have to login")
- Information about the progress for a particular task, possibly giving it a title as well ("Step 1 of 2: Login method"), and proper instructions for what has currently to be done, ("Choose how to login") and what the requirements are ("You will need ...")
- Easy navigation within the process in order to give the user the possibility to go to arbitrary parts of it and to acquire information herself, for example by means of navigation buttons, tabs, or a breadcrumb trail
- Sufficiently short play-back of timed media, depending on the content

### E. Visual comprehension

Some cognitive impairments cause difficulties in processing visual information. This demands for the following non-exclusive list of design principles.

- Use of several modalities to convey a particular message, leaving it to the user to choose the form that best fits her needs. Typical examples of modality sets are {text, still image/graphic} and {text, audio/voice}
- Complementing of still images with offering animations or video, and vice versa, i.e., offering the modality sets {text, graphic, image animation} and {text, still image, slide show, video}
- Voice accompanying a video for improved understanding ability, as voice accompanying the visuals in a video decreases the time needed for understanding
- Presentation of video content (and accompanying voice) in a sufficiently calm manner to give users the time to process the information given
- Screencasts showing a small region of interest instead of the entire screen to allow users to differentiate between the video and the actual page

For a discussion of instructions videos, it is referred to Section IX.

## VII. Benefits for other user groups

Despite the fact that measures for different intellectual deficits sometimes have to be balanced against each other, in general not only a particular target group benefits from certain design principles but rather the vast majority of users.

For instance, the set of design principles for an individual with visual impairments greatly overlaps with the sets for those with language and text difficulties and for those with visual comprehension deficiencies. Next, the group of computer novices shares a considerable number of design measures with users known to have learning and problem solving deficits. Computer novices are likely to have a limited amount of web skills for problem solving and at the same time are untrained regarding the specific problem. This leads to situations where a user is overwhelmed by the technological challenge and therefore lacks the ability to keep the overview, and to focus on what is important in the process.

Tired users typically lack the ability to concentrate over longer periods of time, so they can be categorized as having attention span impairments. Another example is elderly people who sometimes suffer from loss of short-term memory and may have poor concentration skills. This is a compound functional problem consisting of memory and attention span deficits. Finally, even expert users and people with good web skills may be facing challenges when they — being in a new situation — are untrained for a particular task.

To sum up, the majority of design principles for a user group with intellectual impairments helps other user groups as well, and they often are useful for almost any user. This result is consistent with other recent research [25]. After all, a user's cognitive abilities vary over time and typically depend on a particular situation. For instance, consider the situation of a car driver who must not let the eyes off the traffic. Any text message must thus be read out loud to him, which corresponds to the impairment blindness.

## VIII. Prototype

The final login prototype comprises a number of pages, including pages covering various login methods, a new portal page, a personalized "My Page", and a page demonstrating screencast technology. A screenshot of the first step during the login process is shown in Figure 3.

All pages and all page elements were checked (manually) throughout the design process against all parts of the design guidelines to ensure the inclusive result. And as already mentioned, there were several iterative cycles in which the requirements specification and consequently the page design were improved by the feedback from the evaluation phases.

It can be seen that all points of criticism, as expressed in Section IV, have been addressed.

1) Only relevant content is shown.
2) The list of options has to been cut down from 7 options to 4, while the remaining alternatives are "hidden" in a drop-down element
3) The security icon has been removed entirely; instead, a link to more exhaustive security information is provided (in the right-hand "assistance" box). Also, the status bar (on top of the page) displays security information, such as 'insecure connection'
4) The login process has been split up into 2 steps/pages (of which only the first is shown in Figure 3); one where the login method has to be chosen, and another one with the main login part (which depends on the choice in step 1)

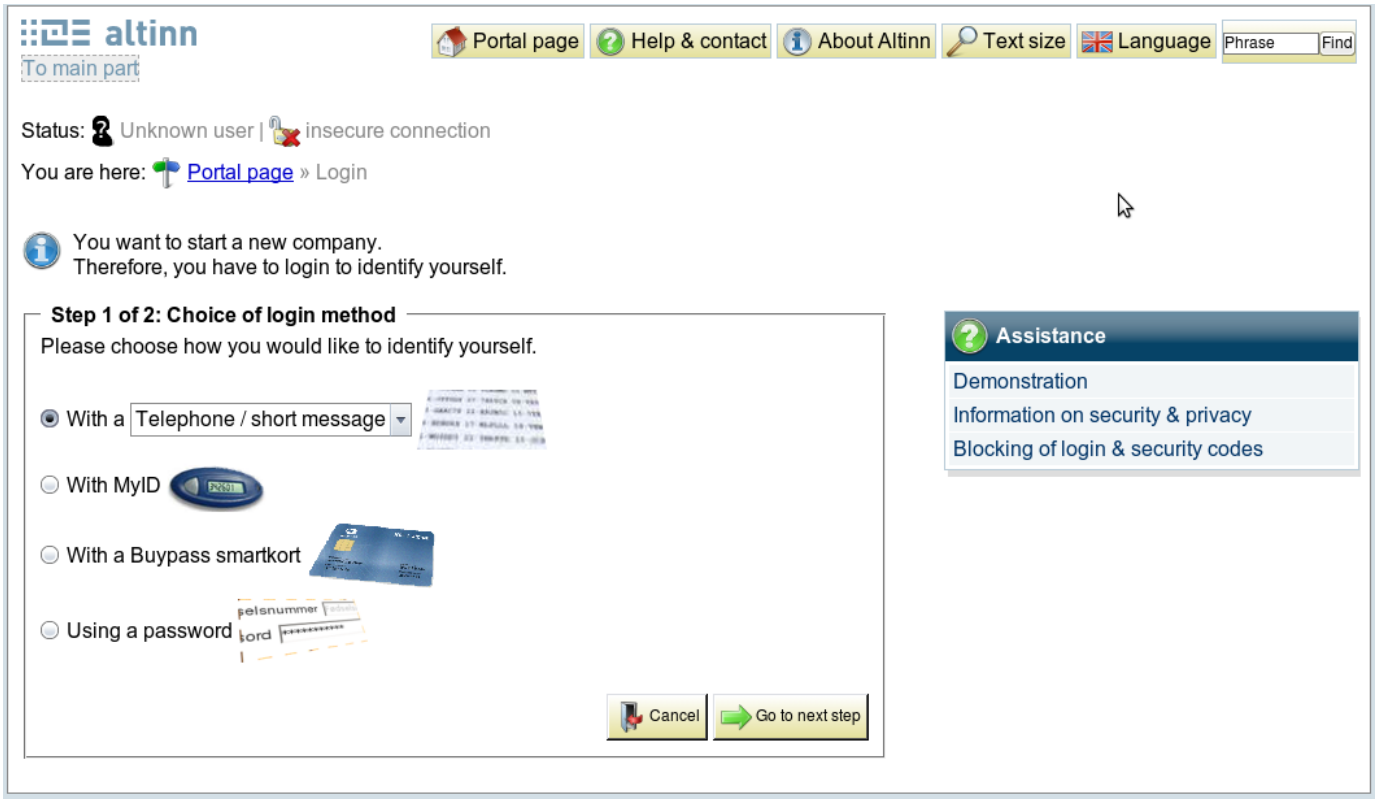



Figure 3. Screenshot of the prototype's first step out of two of the login process

The example illustrates also a number of other design principles as well, such as the use of symbols/icons for fast comprehension, buttons for quick navigation, easily separable blocks of content, etc. User with orientation difficulties in addition to low vision will benefit from a fluid/flexible page layout where the page elements stay approximately in place compared to each other when zooming into the page, i.e., when increasing font and image dimensions. This is also an advantage when the user has a screen with only small dimensions available. Figure 4 shows a combination of both implications.

The prototype also includes a user-specific personalized page which matches the "My page" of today's solution as mentioned in Section IV. A screenshot of this page is given in Figure 5. In the prototype, we have addressed all issues encountered previously, including the following.

1) Now there is a "clean" and simple page structure with blocks of content which are easily distinguishable.
2) All content irrelevant in the given context has been removed. Each content block is summarized by a concise title. Icons are additionally used to convey the meaning of associated text.

## IX. Instruction Videos

During the building of the prototype, there was a high focus on multimodality to improve user interface and experience of the old solution. This involved in particular the use of instruction videos or so-called screencasts, where an example user shows and tells how to solve certain problems, e.g., the task of logging in, by means of screen and voice recordings. Various versions of screencasts were tested in several development cycles with hearing impaired / deaf and in particular elderly individuals in a subproject [26], and the design principles found regarding cognitive deficits are the following.

- Subtitles and boxes with help texts not to far off the screen's middle to catch the user's attention, and easily distinguishable from the video content
- Marker / colored area around the cursor to draw the user's focus
- Particular regions of interest marked with for instance red color
- The page with the screencast opening in the same tab/window, and links for navigating back from the screencasts

A number of other principles, e.g., "textual" (Section VI-A) for text and subtitles, "memory" (Section VI-D) for length of the video, and "visual" (Section VI-E) for the region of interest apply as well for screencasts. Other principles found mainly address the needs of people with sensory (visual) deficits. They are listed here for completion purposes.

- Yellow foreground on black background gives the best visibility of text and image objects
- Offering of a variety of voice presentation concerning a

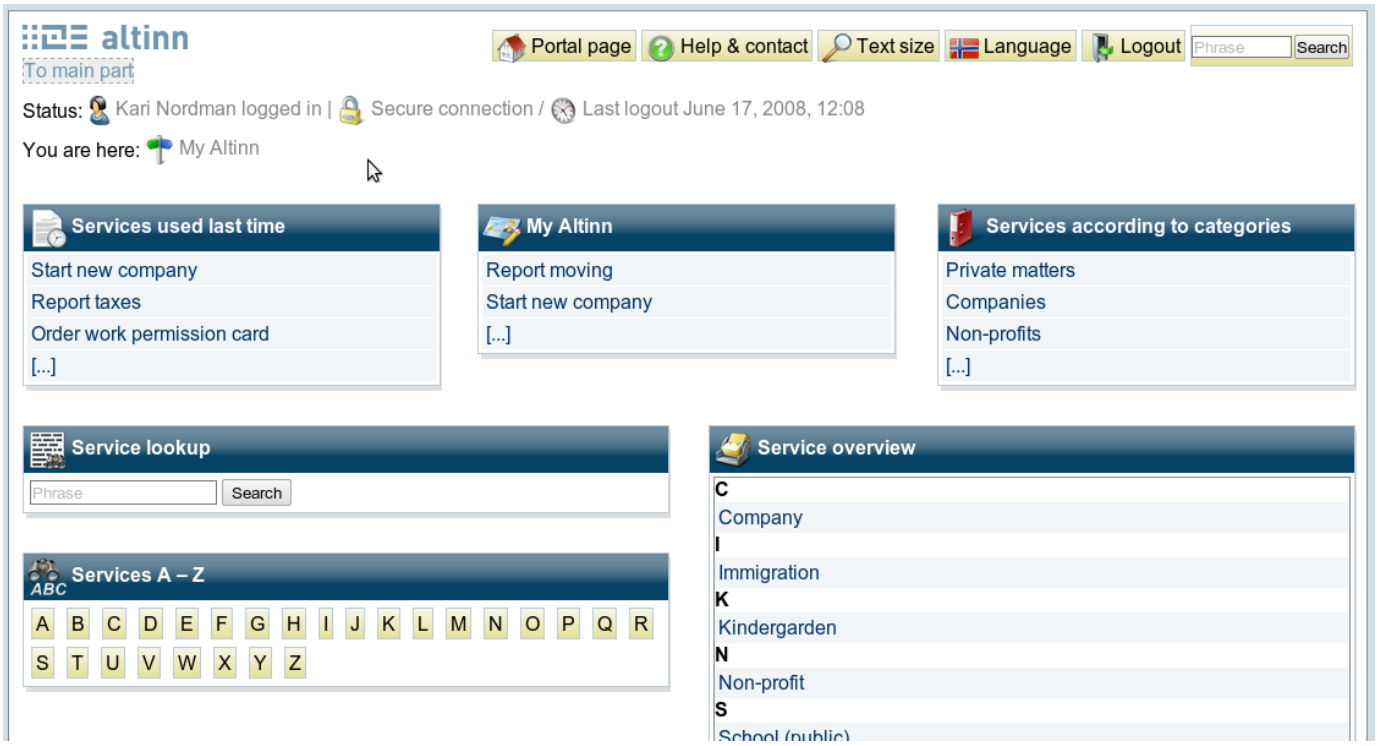



Figure 5. Screenshot of the prototype's personalized "My page"

speaker's gender, pronunciation, and dialect

Another important finding is that the majority of users had problems interacting with the (Flash) media player. I.e., a large number of users was unable to play a video, stop it, replay it if needed, invoke and leave again full-screen mode, etc. One conclusion is that the controls of the media player remain to be hinders in particular for people with cognitive challenges. As the focus of the implementation was on open technologies like international web standards, it has been viewed as outside the scope of the project to modify the proprietary code of Adobe's Flash media player.

## X. Discussion

Concerning the development of static and dynamic web pages, the remedy is to consider intellectual impairments throughout the entire design chain for the system to be built, consisting of

- requirements formulation,
- system architecture and design,
- implementation and integration, and
- testing, evaluation, and verification.

The aim must be to build cognitively accessible solutions, to give users the possibility to participate in the technological progress, and to achieve a stronger market impact of the product or service.

Next, the complexity and heterogeneity of the group labelled cognitively impaired must not be underestimated. As we have seen, cognitive impairments can arise in several ways, and they can affect many aspects of human function. However, some impairments affect only some functions, not others. This has consequences for the design of ICT for these groups. We find that our results correspond with [27] who underscores in a survey of the research in the field, that "one size does not fit all". For example, the principle of simplicity can be recommended on a general level, as we know from other research that most users disapprove complex web pages. However, when applied to specific groups in specific contexts, we will find that a feature or interface solution that one user may find simple, would be too difficult for another. Accordingly, we must avoid defining "simplicity" in general terms and instead understand it in terms of the cognitive skills and capabilities of the actual user and user groups.

Likewise, when it comes to practical interface design, "most users do better with wider interfaces, but some may do better with narrower interfaces" [27]. We must therefore be careful in the formulation of design recommendations to account for the diversity all user groups. This is particularly pertinent for the development of universally designed solutions where there must be an increased focus on utilizing the potential for developing flexible, personalized, and customized solutions. Successful solutions of the future must be adapted to the individual needs and capabilities of each single user. Actual accessibility and usability cannot be reduced to specific features and interface affordances. Our list of recommended design principles is hence not exhaustive and must be applied within an appropriate framework that provides individualization and personalization.

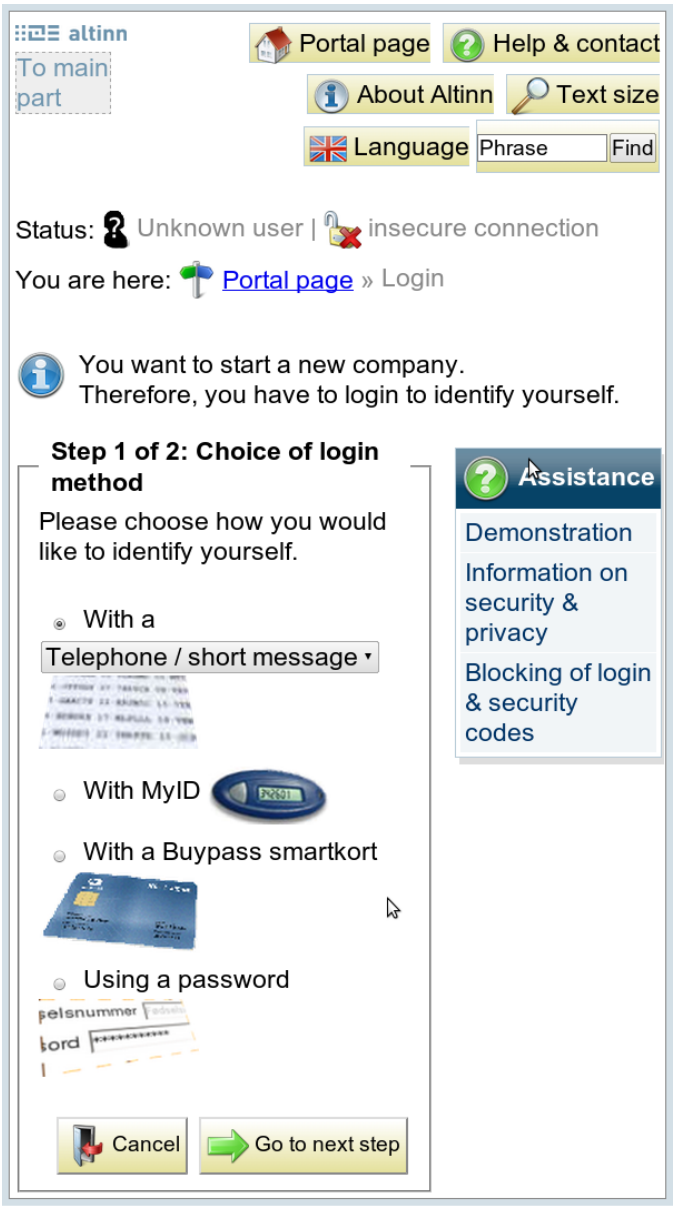



Figure 4. Screenshot of the prototype's login process with narrow page dimensions and a zoom factor of ca. 150%

## XI. Conclusion and Outlook

The login process of an existing website has been made more accessible and usable concerning users with intellectual deficiencies, and in particular with linguistic, learning and problem solving, focus and attention span, memory, and visual-comprehension challenges in mind. Additionally, and in contrast to other work, the topic orientation problems has been addressed.

A number of generic accessibility principles was derived for each deficiency, and these principles were implemented in the improved login solution. This work aims hence at basing the heuristics and educated guesses typically given in the literature on concrete examples. The prototype has undergone several iterations with various testing, including personas, experts, and user feedback. The final testing results show that the prototype provides a solution which suits the needs of the target group much better than today's solution.

Concerning future work, future international standards/recommendations should reflect the knowledge about cognitive deficits and technical remedies regarding static and dynamic web pages in their technical recommendations.

## Acknowledgment

The authors wish to express their gratitude to Riitta Hellman and Gro Marit Rødevand, both with Karde AS, Norway, for their contributions in all associated projects.

## References

[1] T. Halbach, "Towards cognitively accessible web pages," in *Proceedings of International Conferences on Advances in Computer-Human Interactions (ACHI)*. St. Maarten (Netherlands Antilles): IARIA, Feb. 2010.

[2] D. Braddock, R. Hemp, and M. Rizzolo, *The state of the states in developmental disabilities: 2008*. Washington (DC, USA): American Association Intellectual and Developmental Disabilities, 2008.

[3] M. R. Lightner and D. Erdogmus, "Signal processing challenges in cognitive assistive technology," *IEEE Signal Processing Magazine*, vol. 25, no. 5, pp. 103–108, Sep. 2008.

[4] Altinn, "Altinn — Simplified electronic dialogue," 2009, last accessed: 2011-01-20. [Online]. Available: http://altinn.no/

[5] Norwegian Computing Center, Apr. 2009, last accessed: 2011-01-20. [Online]. Available: http://www.nr.no/pages/dart/project_flyer_unimod

[6] L. Udjus, "Gjør døren høy — gjør porten vid," *Stat & Styring*, vol. 1, no. 3, pp. 20–22, 2007, last accessed: 2011-01-20; in Norwegian. [Online]. Available: http://www.karde.no/Lasses%20artikkel%20i%20Stat%20og%20Styring%202007.pdf

[7] WebAIM, "Cognitive disabilities part 2: Conceptualizing design considerations," 2010, last accessed: 2011-01-20. [Online]. Available: http://webaim.org/articles/cognitive/design/

[8] L. Lines, Y. Patel, and K. Hone, "Online form design: Older adults' access to housing and welfare services," in *HCI and the Older Population Workshop*, Leeds (UK), Sep. 2004, pp. 21–22.

[9] K. Pernice and J. Nielsen, "Beyond ALT text: Making the web easy to use for users with disabilities," Nielsen Norman Group, Tech. Rep., 2001, last accessed: 2011-01-20. [Online]. Available: http://www.nngroup.com/reports/accessibility/beyond_ALT_text.pdf

[10] M. Laff and M. Rissenberg, "Cognitive ability measures for accessible web content," *Universal Acess in Human Computer Interaction. Coping with Diversity*, vol. 4554/2007, pp. 722–730, 2007, last accessed: 2011-01-20. [Online]. Available: http://www.springerlink.com/content/k834018137t41705/fulltext.pdf

[11] eGovernment Resource Centre, "Creating sites accessible to people with cognitive disabilities," 2009, last accessed: 2011-01-20. [Online]. Available: http://www.egov.vic.gov.au/index.php?env=-innews/detail:m2754-1-1-8-s-0:n-1366-1-0--

[12] Access Computing, "How can web pages be made accessible to individuals who have cognitive disabilities?" 2009, last accessed: 2011-01-20. [Online]. Available: http://www.washington.edu/accesscomputing/articles?358

[13] WebAIM, "Cognitive disabilities part 1: We still know too little, and we do even less," 2009, last accessed: 2011-01-20. [Online]. Available: http://www.webaim.org/articles/cognitive/cognitive_too_little/

[14] R. Hudson, R. Weakley, and P. Firminger, "An accessibility frontier: Cognitive disabilities and learning difficulties," 2009, last accessed: 2011-01-20. [Online]. Available: http://www.usability.com.au/resources/cognitive.cfm

[15] ——, "Developing sites for users with cognitive disabilities and learning difficulties," 2009, last accessed: 2011-01-20. [Online]. Available: http://juicystudio.com/article/cognitive-impairment.php

[16] *Web Content Accessibility Guidelines 1.0*, World Wide Web Consortium (W3C) Std., May 1999, last accessed: 2011-01-20. [Online]. Available: http://www.w3.org/TR/WCAG10/

[17] *Web Content Accessibility Guidelines (WCAG) 2.0*, World Wide Web Consortium (W3C) Std., Dec. 2008, last accessed: 2011-01-20. [Online]. Available: http://www.w3.org/TR/WCAG20/

[18] L. Seeman *et al.*, "Formal objection to wcag 2.0," 2006, last accessed: 2011-01-20. [Online]. Available: http://lists.w3.org/Archives/Public/w3c-wai-gl/2006AprJun/0368.html

[19] T. Halbach, R. Hellman, G. M. Rødevand, and I. Solheim, "Utformingsveileder for kognitiv tilgjengelighet av elektroniske tjenester og innhold," Feb. 2010, last accessed: 2011-01-20. [Online]. Available: http://iktforalle.no/

[20] E. Gabrielsen, "IALS/SIALS-prosjektet," *SLF-info*, Jun. 2000, in Norwegian.

[21] WebAIM, "Cognitive disabilities," 2009, last accessed: 2011-01-20. [Online]. Available: http://www.webaim.org/articles/cognitive/

[22] A. D. Shaikh, "The effects of line length on reading online news," *Usability News*, vol. 7, no. 2, Jul. 2005, last accessed: 2011-01-20. [Online]. Available: http://psychology.wichita.edu/surl/usabilitynews/72/LineLength.asp

[23] R. Dhamija and L. Dusseault, "The seven flaws of identity management: Usability and security challenges," *IEEE Security & Privacy*, vol. 6, no. 2, pp. 24–29, 2008.

[24] T. Leithead, "Handling script errors from three different perspectives," *The Windows Internet Explorer Weblog*, Apr. 2009, last accessed: 2011-01-20.

[25] W. Quesenbery, "Usable accessibility: Making web sites work well for people with disabilities," Feb. 2009, last accessed: 2011-01-20. [Online]. Available: http://www.uxmatters.com/mt/archives/2009/02/usable-accessibility-making-web-sites-work-well-for-people-with-disabilities.php

[26] I. Haugan, T. Olsen, and P. J. Blakstad, "Multimedial opplæring," Master's thesis, Høgskolen i Oslo, May 2009.

[27] C. Lewis, "Simplicity in cognitive assistive technology: A framework and agenda for research," *Universal Access in the Information Society*, vol. 5, no. 4, pp. 351–361, 2006, last accessed: 2011-01-20. [Online]. Available: http://www.springerlink.com/content/1615-5289/5/4/

# Policies and Abductive Logic: An Approach to Diagnosis in Autonomic Management

Michael Tighe, Michael Bauer
*Department of Computer Science*
*The University of Western Ontario*
*London, ON, N6A 5B7, CANADA*
*Email:{mtighe2;bauer}@csd.uwo.ca*

***Abstract*—Policy-based Autonomic Management monitors a system and its applications and tweaks performance parameters in real-time based on a set of governing policies. A policy specifies a set of conditions under which one or more of a set of actions are to be performed. It is very common that multiple policies' conditions are met simultaneously, each advocating many actions. Deciding which action to perform is a non-trivial task. We propose a method of diagnosing the system to try to determine the best action or actions to perform in a given situation using Abductive Inference. We develop an original method of building a causality graph to facilitate diagnosis directly from a set of policies. We propose two alternate methods of ranking diagnosis hypotheses based on their likelihood of success. Performance of the diagnosis method is evaluated within an autonomic management system monitoring the performance of a LAMP (Linux, Apache, MySQL, PHP) server being governed by the manager. The performance of the diagnosis method is compared against previous methods used by an existing autonomic manager. The results are favourable when compared to previous methods of action selection and to the server running without the autonomic manager. A walkthrough of an example experiment using diagnosis is presented to gain additional insight into the method.**

***Keywords*-Autonomic Computing; Policy; Policy-based Management; Diagnosis; Adbuction**

## I. Introduction

Autonomic Computing represents an effort to make distributed, highly interconnected and interdependent systems into self-reliant systems, capable of configuring, optimizing, healing and protecting themselves [1]. Taking a naming cue from the human autonomic nervous system, the motivation behind autonomic computing is to relieve the massive strain on human Information Technology workers from managing and configuring large systems. The task of installing, configuring, and micro-managing these systems needs to be passed on to the system itself, leaving only high level goals and objectives to be specified by human operators.

Policy-based Autonomic Management aims to fill one piece of the Autonomic Computing vision, by automating the configuration and optimization of several applications running together, in real-time. Performance metrics are monitored for running applications, and configuration parameters are tweaked in real-time to match the current environment and workload [2]. The goal is to achieve some Quality of Service (QoS) objectives [2]. The knowledge used to decide what to change and when to change it is stored within a set of *policies*. The primary type of policy in use in current autonomic management systems is the *action* policy (also called *obligation* or *expectation* policies) [3]. An action policy specifies actions for a manager to perform given that a specific set of events or conditions are present [2]. When the conditions of a policy are true and action should be taken, the policy is said to be *violated*.

In a system containing multiple policies governing the behaviour of the autonomic manager, multiple policy violations advocating many different actions are not only inevitable but are in fact commonplace. The violation of multiple policies may be the result of several discrete problems, or a single problem manifesting itself in several locations. Determining which action to perform out of the set of all actions available is a non-trivial decision [4]. Current work on selecting an action in such a situation has attempted to assign weights to actions based on a number of factors, and then execute the action with the highest weight. We propose to use abductive diagnosis [5] to try and determine the best action to perform. Abductive diagnosis uses knowledge of causal relationships between problems and causes to hypothesize about the specific cause or causes of a given set of problems [6]. This knowledge can be modeled in a bipartite graph. We introduce a method of building such a graph using the policies themselves, with no modifications or other input required. We then test this method by implementing it in an existing Autonomic Management tool.

The remainder of this paper is organized as follows: In Section II we examine related work in Policy-based Autonomic Management and Diagnosis. In Section III, we discuss an Autonomic Management tool, called BEAT, in which we have implemented our diagnosis work. Section IV describes the current method of policy action selection. In Section V, we introduce Abductive Reasoning and Diagnosis. In Section VI, we propose a method of applying Abductive Diagnosis to Policy-based Autonomic Management. In Section VII we describe the implementation of the method as well as our experiments, and present some results in Section VIII. Section IX presents an illustrative example of the diagnosis method in action. Finally, Section X provides conclusions and some thoughts on future work.

## II. Related Work

Work in Policy-based Autonomic Management focuses on both the manager and the policy language itself. One such language is Ponder [7], an object-oriented, declarative policy language. It is designed as a generic language that can be used in a number of different implementations [7]. AGILE [8] is another policy language, designed for flexibility and to provide run-time adaptation of policies. It has been developed as part of a larger policy-based autonomic management system. Lymberopoulos et al [2] have developed a policy-based framework in which policy adaptation is the key focus. The authors construct a framework for network services management, with policies capable of being dynamically adapted to meet a changing workload and environment [2]. Other policy languages include PDL (Policy Description Language) from Bell Labs [9] and CIM-SPL (Common Information Model Simplified Policy Language), which is from the DMTF (Distributed Management Task Force) and is part of their larger CIM [10].

Abductive logic has been used in policy conflict detection by Bandara et al [11]. In this method, conflict detection is done prior to execution on a formalized version of the policies. Other work on diagnosis in autonomic computing has developed outside of policies, and has focused on determining the component that is the root cause of a given problem using machine learning methods. Duan and Babu [12] developed a system called *Fa* which monitors a large number of system metrics and performs diagnosis using a learned classifier. The classifier is learned via supervised learning by annotating sets of system metric values with failure states. Ghanbari and Amza [13] combine models for anomaly detection on individual components together into a single belief network, modelling the structure and causal relationships of components. Learning based on observing injected faults is performed to refine the model and the probabilities associated with the causal relationships. Also, the individual models used to detect component anomalies must be trained.

Previous work on diagnosis has not considered the use of policies. Our approach to diagnosis therefore differs from previous work in that it focuses specifically on diagnosis in policy-based autonomic management. We make use of the structure and content of the policies themselves, thus making our approach domain independent.

## III. BEAT Autonomic Manager

The diagnosis algorithm has been implemented in a previously developed Autonomic Management tool, called BEAT (Best Effort Autonomic Tool). BEAT is a policy-based autonomic management framework, described in [3], [4]. Policies are used to specify how the management is performed as well as how the manager itself operates. The BEAT management system knows how to monitor and manipulate the system, and the policies provide the necessary rules to dictate how such manipulations should be carried out. These are low-level policies specifying specific actions to be taken under specific circumstances. Individual monitored system and application metrics are used to determine the situation (conditions) and actions consist of the modification of specific tuning parameters.

```
expectation policy RESPONSETIMEViolation
if (APACHE:responseTime > 2000.0) &
(APACHE:responseTimeTREND > 0.0) then
   AdjustMaxClients(+25)
     test MaxClients + 25 < 501 |
   AdjustMaxKeepAliveRequests(-30)
     test MaxKeepAliveRequests - 30 > 1 |
   AdjustMaxBandwidth(-128)
     test MaxBandwidth - 128 > 128
end if
```

Figure 1: Pseudo-code Response Time Policy

A single policy, or policy rule, consists of two main components: A set of conditions, and a set of actions. Figure 1 shows a pseudo-code version of a typical policy in BEAT. This policy describes what should occur when the response time of a web server exceeds a certain threshold. Note that this is only a representation of a policy. Actual policies in BEAT are not written in this way, and are instead built in a GUI and stored within a relational database. The policy essentially states that given that these conditions hold true, one of these actions should be performed (if CONDITIONS then ACTIONS).

Policy conditions compare some system metric to a value using a specified operator, and can be combined using standard logical operators. A single condition is not unique to one policy, but can be contained within several policies within the system. In Figure 1, the conditions are APACHE:responseTime $>$ 2000.0 and APACHE:responseTimeTREND $>$ 0.0. In these cases, the metrics being monitored are the response time of the Apache web server and the recent trend of the response time.

Policy actions specify some system or application parameter to be modified in response to the violation of the policy. For example, in Figure 1, AdjustMaxClients(+25) is an action modifying the Max Clients parameter of Apache by increasing it by 25. The action may also contain a test that must be performed and passed before execution. This could be used, for example, to prevent modifying a value beyond some hard upper and lower bounds. The AdjustMaxClients(+25) action is associated with the test MaxClients + 25 $<$ 501, thus enforcing an upper bound of 500 on the Max Clients parameter. Again, a single action may be advocated by multiple policies. In addition, a policy will specify a list of actions, with the implication that only one should be executed, but not all. The decision as to which action to execute falls on the Autonomic Manager itself.

The BEAT Autonomic Manager consists of several com-

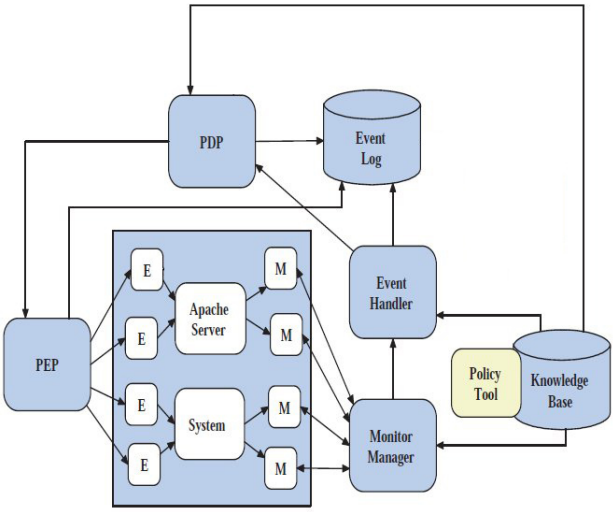



Figure 2: BEAT Autonomic Manager Architecture [3]

ponents which interact with each other to provide the full management functionality. Figure 2 [3] shows the general architecture of the system. *Monitor* components (labelled *M*) monitor the state of the system and running applications being managed, and forward this information to the *Monitor Manager*. The Monitor Manager aggregates and processes this information, and generates events which are sent to the *Event Handler*. This component then determines if the events are of interest (if they represent a violation of a policy), and forwards events to the *Policy Decision Point (PDP)*. The PDP uses the policy information to determine what, if any, action should be taken. Actions to be executed are then sent to the *Policy Enforcement Point (PEP)*, which is responsible for executing the action. The PEP determines if the action can be performed, and if so, forwards it to an appropriate *Effector* component (labelled *E*), which performs the actual modification to system and application parameters. Policies and other information are stored in the *Knowledge Base* and manipulated via a *Policy Tool*. The *Event Log* records previous events.

## IV. Policy Action Selection

In a system containing multiple policies governing the behaviour of the autonomic manager, multiple policy violations advocating many different actions are not only inevitable but are in fact commonplace. The violation of multiple policies may be the result of several discrete problems, or a single problem manifesting itself in several locations. Determining which action to perform out of the set of all actions available is a non-trivial decision [4]. There are a few ways in which an action can be selected. One possibility is to simply select the first action that arises, which is essentially an arbitrary selection. This method takes nothing into account in its decision making, leaning heavily on the expertise of the policy designer, and therefore seems to be a poor choice. Another option is to weight policies and actions based on some criteria. Possible criteria include [4]:

- The *severity* of the violation, which refers to how far a threshold value on a metric has been exceeded.
- Manually assigned weights on policy conditions.
- The *advocacy* of the action, which refers to the number of violated policies advocating the same action.
- The *specificity* of the policy, which refers to the number of conditions used to trigger the policy, assuming that policies containing more conditions should be dealt with first.

These criteria and others can be used separately or in combination to provide some guidance in the action selection process. These criteria are based on intuition, and it is unclear how well they choose the best action to execute. Another possibility is to employ machine learning techniques to learn the "best" action to select in a given circumstance, based on previous experience [14], [15]. Again, this could be used in conjunction with other techniques to improve the action selection mechanism.

If an incorrect action is selected and taken, not only is time wasted before the correct action can be selected, but the modification of application tuning parameters that should not have been modified may cause further problems. This makes action selection a key problem in the performance of an autonomic manager.

## V. Abduction and Diagnosis

Abductive reasoning is an alternative to deductive and inductive reasoning. This form of reasoning most closely resembles how a human diagnoses problems. Let us say that a problem consists of a set of rules, a specific case, and a result that occurs given the two. In abductive reasoning, we have the set of rules and the result, and we hypothesize about the specific case that is causing the result [6]. For example, if a doctor is diagnosing a patient, the set of symptoms experienced by the patient would be analogous to the result and the doctor's medical knowledge would be the set of rules. The doctor's diagnosis as to what the potential ailments the patient could have would be the set of specific case hypotheses. Note that unlike deduction and induction, we do not arrive at a definitive decision or conclusion; we can only build hypotheses representing what the specific case might be [6].

Peng and Reggia [6] present a formal method of representing and diagnosing an abductive reasoning problem. Given a set of *disorders* representing underlying problems, a set of *manifestations* representing observable symptoms, and knowledge of the causal relationship between the two, adbuctive methods can be used to build a diagnosis. This can be represented by a graph, which we will call a *Causal Network*, containing both the disorder and manifestation sets. An edge from a disorder to a manifestation indicates that the disorder may cause the manifestation, although

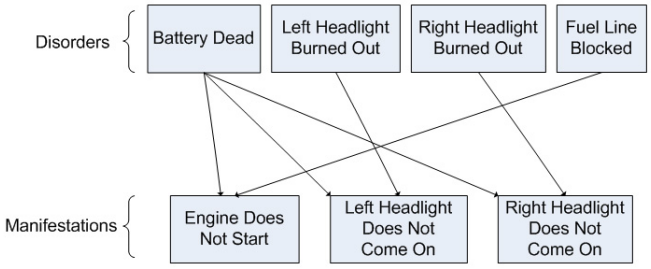



Figure 3: Causal Network Example (based on example from Peng and Reggia [6])

it is important to note that it may not. A disorder can cause multiple manifestations, and a manifestation may be caused by many different disorders. Given a set of currently present manifestations and the causal network, diagnosis can be performed to build a set of hypothesis disorder sets that could explain the manifestations. It is impossible to guarantee that a definitive diagnosis can be obtained. The best that can be achieved is the construction of a set of hypotheses. Each hypothesis contains a set of disorders that fully explain the present manifestations, but determining which hypothesis is correct or even which hypotheses are more likely to be correct is a non-trivial task.

A simple example is given in Peng and Reggia [6] describing a causal network for the diagnosis of automotive problems. It uses a small set of disorders and manifestations and presents the causal associations between them that form the causal network graph. The disorders include *battery dead*, *left headlight burned out*, *right headlight burned out*, and *fuel line blocked*. The manifestations are *engine does not start*, *left headlight does not come on*, and *right headlight does not come on*. Figure 3 shows the causal network for these disorders and manifestations, including the causal associations between them.

A *cover* of a given set of present manifestations is a set of disorders such that each present manifestation can be caused by at least one disorder in the set. Each cover represents a single hypothesis solution, giving one potential explanation for the manifestations. Peng and Reggia [16] suggest that simpler covers are more likely to be true than complex ones. It is then these simple covers that we wish to find when diagnosing a problem. There are several different suggested criteria for judging the simplicity of a cover. A *single-disorder* cover is a cover that consists of only a single disorder. An *minimal* cover contains the minimal number of disorders required to cover all present manifestations. An *irredundant* cover is a cover where each disorder causes at least one manifestation that no other disorder in the cover causes. A *relevant* cover is a cover that contains no disorders that are not a cause of at least one present manifestation. These criteria create increasingly broad sets of covers as we move from single-disorder to relevant covers. The set of single-disorder covers for a set of manifestations is a subset of the set of minimal covers, which is a subset of the set of irredundant covers, which is finally a subset of the set of relevant covers.

Of these criteria, irredundancy seems intuitively to be the best choice. Single-disorder covers are unnecessarily restrictive, and clearly insufficient in situations where more than one problem (disorder) can occur simultaneously. Minimal covers are also too restrictive, as it is easy to imagine a case where a minimal cover is clearly not the most likely explanation of the manifestations. For example, in medical diagnosis, a minimal cover may consist of a single rare disease, where another cover may exist containing two common diseases. Clearly the minimal cover is less likely in this case. Relevant covers, on the other hand, represent the other extreme in which far too many covers are accepted as plausible. Irredundant covers will therefore be used for diagnosis.

An algorithm has been developed by Peng and Reggia in [6] to diagnose the problem by constructing the set of all irredundant covers of the present manifestations. The actual diagnosis algorithm, presented in Figure 4 is quite simple. The algorithm starts with a set of hypotheses, and is given the set of current manifestations. It then iterates through each manifestation (in no particular order), revising the hypothesis set each time to represent all irredundant covers of the new manifestation as well as previously added manifestations. Once all manifestations have been accounted for, the algorithm is complete.

```
1: hypothesisSet = {∅}
2: while moreManifestations do
3:    m_new = nextManifestation;
4:    hypothesisSet = revise(hypothesisSet, causes(m_new))
5: end while
6: return hypothesisSet
```

Figure 4: Diagnosis Algorithm

The heart of the algorithm is the *revise* method. The method, which accepts the current set of hypotheses as well as the set of causes for the manifestation being added, results in a new hypothesis set that contains all irredundant covers of the set of manifestations that have been processed, including the new one. It does this by first finding all hypotheses in hypothesisSet that are also irredundant covers of the new manifestation, and leaving them unchanged. It then modifies any remaining covers so that they also cover the new manifestation by adding new disorders to them. Finally, any duplicate or irredundant covers created in the last step are removed. Details of the algorithm and revise method can be found in [6].

Let us return to our basic example of automotive diagnosis from [6], illustrated in Figure 3, and take a high level look at the diagnosis algorithm in action. Let us say that we

are currently observing all of the possible manifestations, that is, *engine does not start*, *left headlight does not come on*, and *right headlight does not come on*. We start with an empty hypothesis set in line 1 of the algorithm, and lines 2 to 5 proceed to loop through each presently observed manifestation, one at a time. The first manifestation, *engine does not start*, is retrieved in line 3. Line 4 revises our current hypothesis set, which is empty, by including the causes of *engine does not start*, which are *battery dead* and *fuel line blocked*. This results in a set of two hypotheses, namely, either *battery is dead* or *fuel line is blocked*.

Next we add the *left headlight does not come on* manifestation, and revise the hypothesis set with its causes (*battery dead* and *left headlight burned out*). The revise method first picks out *battery is dead* as a cover of both the original manifestation and the new one, and decides to leave it in the hypothesis set. Next, it modifies the second cover, *fuel line is blocked*, so that it covers the new manifestation. This is done by adding the disorder *left headlight burned out*. Other covers are possible, but would not be irredundant. This results in a new set of hypotheses, where either *battery is dead* is true (which would cover both manifestations itself), or *fuel line is blocked* and *left headlight burned out* are both true.

Finally, we add the last manifestation, *right headlight does not come on*, using its causes (*battery dead* and *right headlight burned out*). This brings us to our final set of hypotheses, which explains (covers) all three presently observed manifestations. We still have only two hypotheses. Either the disorder *battery dead* is true, which itself explains all three manifestations, or *fuel line blocked*, *left headlight burned out* and *right headlight burned out* are all true simultaneously. Both of these hypotheses are irredunant covers of the set of presently observed manifestations. Logically, it makes sense that given the manifestations, either the battery is dead or the fuel line is blocked and both headlights are burned out.

## VI. Mapping Policies to a Causal Network

In order to perform diagnosis, a Causal Network containing potential disorders, manifestations, and their relationships must be constructed. We do this based on existing information contained within the set of policies governing the Autonomic Manager, at run-time. In this way, diagnosis can be used for action selection without requiring any extra diagnosis-specific information to be added. This method of bridging the gap between policies and abductive diagnosis constitutes our main original contribution.

Each policy contains a set of conditions and a set of actions. These conditions and actions are not unique, and the same conditions and actions will be used by many different policies. The manifestations can be derived from the conditions, in that each condition is directly mapped to a single manifestation. If the condition is true, then the equivalent manifestation is considered present. Disorders are derived from the policy actions, with each action being used to build a single disorder. Since an action is intended to correct some parameter that is thought to be set incorrectly for the current environment and workload, then that parameter being incorrectly set can be considered the underlying disorder causing the manifestations. For example, if an action specifies that the Max Clients parameter of the Apache server should be increased by 25, then the disorder derived from such an action would be *Apache Max Clients too low*.

Associations between the generated manifestations and disorders can be easily derived from the policies as well. If a policy containing the condition used to derive a certain manifestation also advocates the action used to derive a certain disorder, then that manifestation could potentially be caused by the disorder and should be associated with it. Since conditions and actions are replicated across many different policies, this method results in a fairly well connected Causal Network. Diagnosis can then be performed using this Causal Network, essentially finding the action or actions that can potentially cause all present conditions to no longer be true, thus eliminating all policy violations. Figure 5 gives a psuedo-code version of the algorithm.

```
1:  disorderSet = {∅}
2:  manifestationSet = {∅}
3:  causalRelationships = {∅}
4:  for all policies p do
5:     for all conditions c in p do
6:        m = Manifestation(c)
7:        if m not in manifestationsSet then
8:           manifestationSet += m
9:        end if
10:       for all actions a in p do
11:          d = Disorder(a)
12:          if d not in disorderSet then
13:             disorderSet += d
14:          end if
15:          causalRelationships += (d, m)
16:       end for
17:    end for
18: end for
19: return  hypothesisSet
```

Figure 5: Policy Mapping Algorithm

The size of the Causal Network depends on the number of policies deployed in the system, and the level of redundancy in the policy conditions and actions. To look at one extreme, if each policy contains completely unique conditions and actions, it will result in a graph with many nodes and very few connections. If, on the other hand, the conditions and actions are replicated throughout many different policies (which is the usual case), the graph will have fewer nodes

and be well connected. Even with a large number of policies, the Causal Network will remain a simple bipartite graph, and should scale well as the number of policies increases.

```
expectation policy RESPONSETIMEViolation
if (APACHE:responseTime > 2000.0) &
(APACHE:responseTimeTREND > 0.0) then
   AdjustMaxClients(+25)
     test MaxClients + 25 < 501 |
   AdjustMaxKeepAliveRequests(-30)
     test MaxKeepAliveRequests - 30 > 1 |
   AdjustMaxBandwidth(-128)
     test MaxBandwidth - 128 > 128
end if

expectation policy CPUViolation
if (CPU:utilization > 85.0) &
(CPU:utilizationTREND > 0.0) then
   AdjustMaxKeepAliveRequests(-30)
     test MaxKeepAliveRequests - 30 > 1 |
   AdjustMaxBandwidth(-128)
     test MaxBandwidth - 128 > 128
end if
```

Figure 6: Pseudo-code CPU Policy

Figure 7 shows an example Causal Network derived from example policies in Figure 6. The CPUViolation specifies actions to be taken to lower CPU utilization in the event that it exceeds a threshold value of 85%. For simplicity in the diagram, the Manifestation nodes that would be generated for the trend conditions (APACHE:responseTimeTREND and CPU:utilizationTREND in Figure 6) are omitted, as they would be identical to the nodes derived from APACHE:responseTime and CPU:utilization, respectively. In the actual causal network, they would be present.

Since each action must pass an associated test before it can be executed, diagnosis must generate a list of hypotheses, with the first one that passes its associated tests being executed. The set of all hypotheses obtained from diagnosis must therefore be ordered from most likely to be effective to least likely to be effective. The only ranking method currently implemented is to sort the hypotheses based on the number of disorders they contain. This can either be done in ascending or descending order, causing the system to favour either hypotheses containing fewer disorders or more disorders, respectively. This translates to the system either preferring to execute fewer actions when using ascending or preferring to execute many actions with descending.

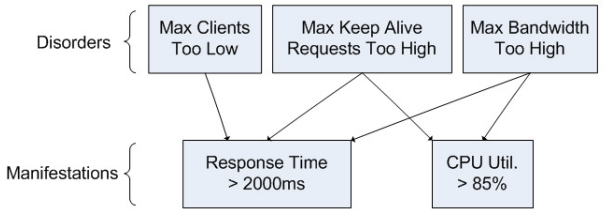



Figure 7: Policy derived Causal Network

Let us see how this causal network could be used for some very basic diagnosis. Say that we observe the manifestation *Response Time* $>$ *2000ms*. We can see that any of the disorders could be the cause, that is, our set of hypotheses includes three single disorder hypotheses (*Max Clients Too Low*, *Max Keep Alive Requests Too High*, and *Max Bandwidth Too High*). Now, let us say that we also observe the manisfestation *CPU Utilization* $>$ *85%*, and update our set of hypotheses using the causes of this manifestation. This reduces our set of hypotheses to two, namely, *Max Keep Alive Requests Too High* and *Max Bandwidth Too High*. You may note that a hypothesis containing both *Max Bandwidth Too High* and *Max Clients Too Low*, or both *Max Keep Alive Requests Too High* and *Max Clients Too Low* would also be a potential cover of the present manifestations. In both cases, however, the disorder *Max Clients Too Low* would be redundant, and we only wish to look for *irredundant* covers, as discussed in Section V.

Hypotheses containing disorders must then be translated back into something useful to the autonomic manager, that is, a list of actions or sets of actions to perform. For each hypothesis, each disorder contained within it is used to look up the original action used to build the disorder. The actions for a single hypothesis are grouped together, and if executed, the entire group must be executed together, since all disorders contained in the hypothesis are required to cover the present manifestations. Using our example, the two single disorder hypotheses of *Max Keep Alive Requests Too High* and *Max Bandwidth Too High* would be translated back into the actions AdjustMaxKeepAliveRequests(-30) and AdjustMaxBandwidth(-128), as seen in Figure 6.

## VII. Experiments

The diagnosis action selection method was implemented in the BEAT Autonomic Manager discussed in Section III. Modifications to BEAT were made in the Policy Decision Point (PDP) component, inserting diagnosis in place of the existing method. We compared the diagnosis method to other methods of selecting policy actions by configuring BEAT to manage a web server, and measuring its performance under a stressful workload. Performance is measured with the autonomic manager using the action selection method previously used in BEAT (describe in Section IV), with two variations of the newly developed diagnosis algorithm, and with the server running without intervention by the manager. Policies are specified with the goal of maintaining specific response time, CPU utilization, and memory utilization ranges, and the methods of action selection can be compared on how well they achieve these objectives. Service differentiation will be used and controlled by the autonomic manager. Incoming requests to the server are divided into three service classes, namely, gold, silver and bronze, with gold being given highest priority and bronze lowest.

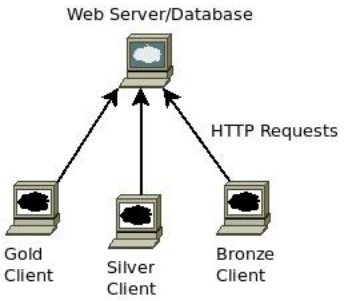



Figure 8: Experimental Setup

### A. Test Environment

Figure 8 shows the basic setup of our experiments.

*Server:* The server machine hosts a web server running a PHP bulletin board application [17]. The application makes use of a database, also running on the server machine. The server is a LAMP stack (Linux, Apache, MySQL, and PHP), with the BEAT Autonomic Manager installed.

*Clients:* There are three client machines responsible for generating the workload for the server. Each machine represents one service class, namely, gold, silver and bronze. Requests sent by the gold machine are to be given highest priority, and requests sent by the bronze machine are to be given the lowest. In a real world implementation, requests could be divided into classes based on a pricing plan, importance as a part of a larger system, or some other prioritization scheme. Load was generated using Apache JMeter 2.3.4 Load Generator [18].

### B. Systems Under Test

Four configurations will be contrasted with each other to evaluate the performance of the diagnosis algorithm. These include the system without the aide of BEAT, with BEAT enabled and using the previously developed action selection method, and finally with two different versions of the diagnosis algorithm.

*Policies Disabled:* A base configuration with the BEAT autonomic manager disabled, and with differentiated services disabled (all requests are treated equally). This will provide a frame of reference for judging the performance improvement offered by the autonomic manager with each form of action selection.

*Weighted Actions:* Section IV outlines a set of criteria that can be used to guide action selection. These criteria, (severity, specificity, weight and advocacy), are implemented in BEAT [15] and combined together to determine a total weight for each action, with higher weighted actions being given priority. Details of this can be found in Bahati et al. [15]. For the purposes of this experiment, we will refer to this as the *Weighted Actions* method.

*Diagnosis with Fewer Disorder Priority (Diagnosis - Fewer):* The first of the two forms of the diagnosis algorithm is Diagnosis with Fewer Disorder Priority. The algorithm itself for both forms is identical. The difference is in the ordering of hypotheses. In this form, hypotheses containing fewer disorders are ranked higher than hypotheses with more disorders. Essentially, this makes the assumption that the simplest hypothesis is most likely the correct one. In many cases this will result in a single action being taken, but it is not necessarily always the case.

*Diagnosis with Many Disorder Priority (Diagnosis - Many):* This second form of the diagnosis algorithm reverses the ordering of the first. It makes the assumption that taking multiple actions will be more likely to be successful than taking a single action. As such, hypotheses containing more disorders will be given priority over those containing fewer. The diagnosis algorithm is otherwise unchanged from Diagnosis with Fewer Disorder Priority. This will often result in multiple actions being taken.

### C. Measures of Performance

Four metrics will be measured to determine the relative performance of each version of the Autonomic Manager.

- *Apache Response Time (Server)* - This is the response time of the Apache web server as measured from the server itself. This value is extremely important, as it is the measure by which the autonomic manager itself determines how well the server is performing. It measures response time by continuously requesting a single page from the web server and measuring the time it takes to receive it. It does not use the KeepAlive option, meaning that a new connection must be opened for each request. It is also independent of the service differentiation mechanism used for requests received from external machines.
- *CPU Utilization* - This is the percentage of the CPU currently in use on the server machine. This does not refer to the amount of CPU being used by the web server only, but rather the total CPU usage.
- *Memory Utilization* - This is the percentage of the total memory that is in use on the server. Like the CPU Utilization metric, this represents total memory usage for the entire server machine.
- *Client-side Response Time* - This is the response time as measured by the client machines. The time taken to complete each request (from the time the request is sent until the entire page has been received) is recorded. These requests make use of the KeepAlive option, meaning that requests sent by a single 'user' attempt to re-use the same existing connection. The set of all client-side request response times can be divided into the three service classes (gold, silver and bronze) to analyze the effects of service differentiation.

### D. Workload

All three client machines ran identical workloads, and were started and stopped simultaneously. The workload was designed to overload the system to a point where without the aid of the autonomic manager, the CPU is running at 100% and server-measured response times are over the 2 second threshold. The workload started with a single thread (or user), and ramped up linearly to a total of 25 threads (or users) over a period of 8 minutes. Each thread continuously performed a small loop consisting of a "think-time" delay of 750-1250ms, and a request to a page randomly chosen from 24 dynamic (PHP generated) pages offered by the PHP Bulletin Board application running on the server. A request included retrieving the HTML page as well as all other resources (images, etc.) contained on the page. This continued for one hour, at which point the test was halted. Each thread used the KeepAlive option, thus attempting to reuse its existing connection to the server as much as possible to avoid reconnecting.

### E. Policy Goals

The policies are designed in such a way as to maintain certain performance objectives, or goals. These typically consist of a threshold value on a measured metric within the system. The duty of the autonomic manager is to achieve these goals as best it can. These goals roughly translate to the conditions of the policies. When the conditions are violated, the policy actions are intended to attempt to push the metric back under the threshold. Without going into detail as to the specific policies and policy actions, the following are the general goals of the set of policies used in this experimentation:

- Apache HTTP server response time, as measured from the server should be below 2 seconds.
- The CPU Utilization should be below 90%. If utilization falls below 85%, then more CPU resources should be used, if needed. Essentially, the system should make use of as much CPU as it can up to the 90% threshold.
- Memory Utilization should be below 50%.
- Priority should be given to the gold, and then the silver and finally the bronze service classes in that order.

## VIII. Results

The experiment was performed identically with each of the four systems under test (Policies Disabled, Weighted Actions, Diagnosis - Fewer, and Diagnosis - Many). Each test was run for exactly one hour, and repeated a total of 5 times. Averages and standard deviations were calculated for each run and averaged over the 5 repeats of the experiment.

Table I shows the metrics measured on both the server and client machines. These include the response time of the Apache web server (as measured by the mechanism described in Section VII-C), CPU Utilization and Memory Utilization. The values shown are the average values for an entire run. The metric averages are then averaged across all 5 replications of the experiment. Figure 9 shows the Apache Response Time data for all experiment replications as a box plot, and figure 10 is a box plot the CPU Utilization for all experiments replications.

| | Disabled | Weighted | Diag. Fewer | Diag. Many |
|---|---|---|---|---|
| Apache Resp. | 3336ms | 1195ms | 1031ms | 1163ms |
| CPU Util. | 98.3% | 74.4% | 82.5% | 82.4% |
| Memory Util. | 22.3% | 24.6% | 24.0% | 24.1% |
| Gold Avg. | 2182ms | 1798ms | 1389ms | 1465ms |
| Silver Avg. | 2228ms | 4021ms | 3920ms | 3827ms |
| Bronze Avg. | 2192ms | 4742ms | 5543ms | 5239ms |

Table I: Average Results

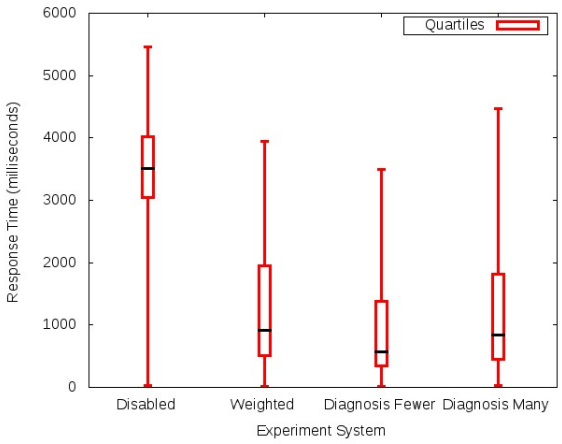



Figure 9: Apache Response Time Box Plot

Judging by the measured response times of the Apache web server, we can easily see that the three tests performed with the autonomic manager outperform the system with the manager disabled. CPU utilization also comes down under the threshold value, while memory utilization increases by a trivial amount and stays well below threshold levels. The response times for the three action selection methods are

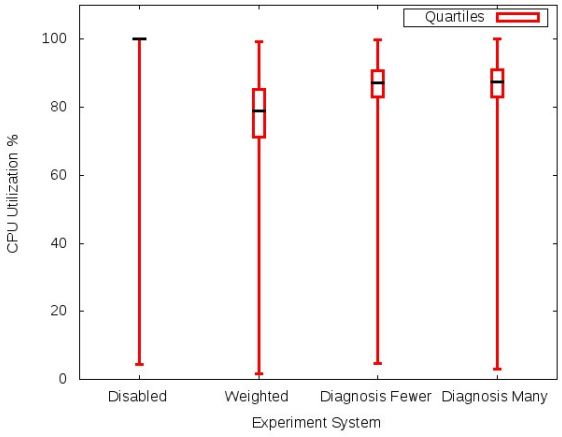



Figure 10: CPU Utilization Box Plot

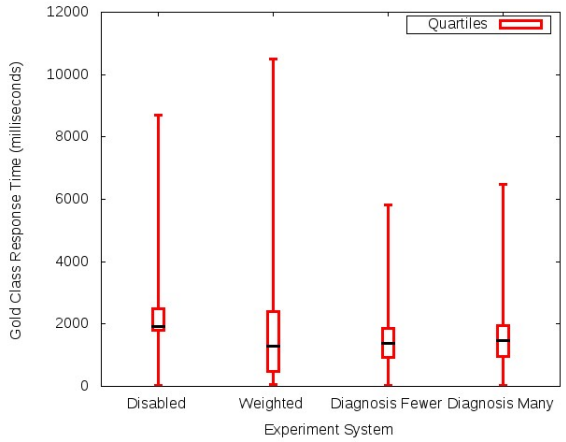



Figure 11: Gold Response Time Box Plot

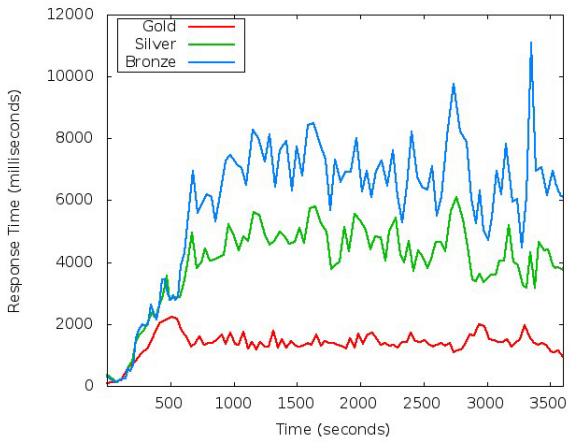



Figure 12: Client Response Times for Diagnosis Fewer

similar, with Diagnosis Fewer (favouring hypotheses with fewer disorders, or fewer actions to take) beating out the other two, which match up fairly evenly. Both variations of the diagnosis algorithm also make better use of the CPU, as outlined in the goals of our set of policies in Section VII-E, without going over the 90% threshold.

The Gold, Silver, and Bronze response times in table I are the client-side response times of the gold, silver, and bronze client machines, respectively. Response times experienced by the client machines show a different side to the performance of the web server than those measured by the server-side response time monitor. As mentioned in Section VII-C, the server-side response time metric does not use KeepAlive, while the client machines do. This means that the server monitor needs to open a new connection for each request, and as such potentially wait in a queue again. Another difference comes from the effect of service differentiation on the client requests. The most important of the client response time measures is that of the gold service class, as the silver and bronze classes should be sacrificed to maintain its performance. The test is designed to put the server under stress, and as such we should see response times of the silver and bronze classes sacrificed to maintain the performance of the gold class. Figure 11 is a box plot of the Gold Class Response Time data for all experiment replications. Both diagnosis algorithms perform better than weighted action selection on gold response time, as well as silver response time, with diagnosis favouring fewer actions having the edge. Figure 12 shows the gold, silver, and bronze response times for the system using the diagnosis algorithm favouring fewer actions, for a single repetition of the experiment. Service differentiation is clearly visible in this graph, as the system keeps the gold class consistent at the expense of silver and bronze.

Figure 13 compares the Apache response times for weighted action selection and diagnosis favouring fewer

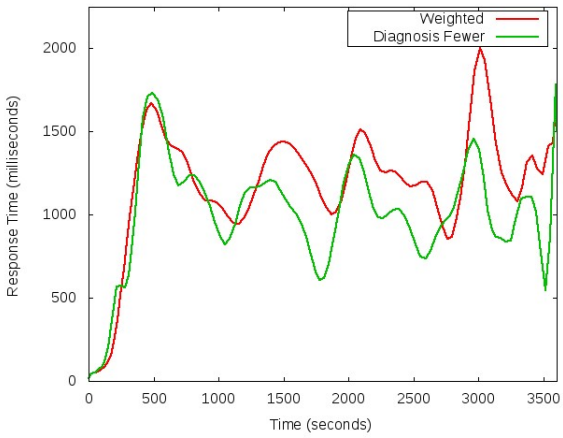



Figure 13: Apache Response Time

actions, for a single repetition of the experiment. The graphed curves are Bezier curve approximations of the actual data, in order to more clearly show the difference between the performance of the two methods. A Bezier curve is a parametric curve approximation of the data used to smooth the data. The data shown is from a single experiment, not averaged over all 5 repetitions, and represents results consistent with all experiments.

Table II shows the same metrics as table I, except only for the overload period of the experiment. That is, the ramp up time to the maximum load of 25 clients per machine (for a total of 75 clients) is excluded, leaving only the time

| | Disabled | Weighted | Diag. Fewer | Diag. Many |
|---|---|---|---|---|
| Apache Resp. | 3722ms | 1300ms | 1095ms | 1247ms |
| CPU Util. | 99.9% | 78.0% | 86.8% | 87.2% |
| Gold Avg. | 2333ms | 1872ms | 1398ms | 1463ms |
| Silver Avg. | 2365ms | 4442ms | 4387ms | 4257ms |
| Bronze Avg. | 2336ms | 5404ms | 6657ms | 6233ms |

Table II: Average Results - Max Load Only

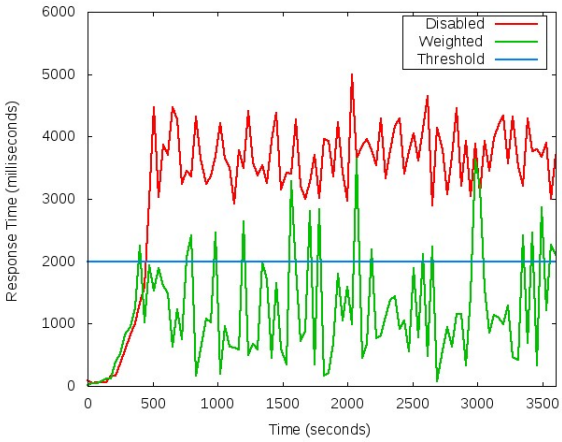



Figure 14: Apache Response Time Over Threshold

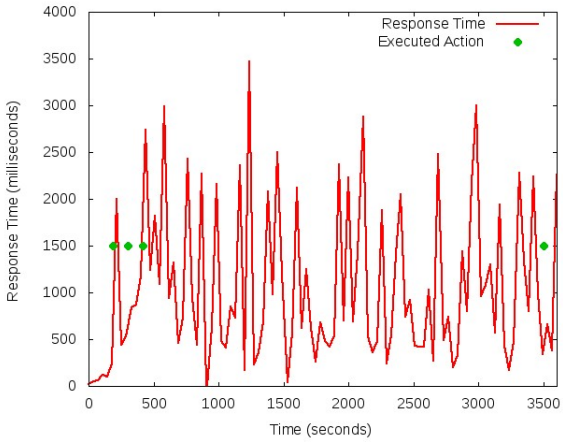



Figure 15: Example Run with Diagnosis - Response Time

| | Disabled | Weighted | Diag. Fewer | Diag. Many |
|---|---|---|---|---|
| Apache Resp. | 290.30 | 14.94 | 11.19 | 16.19 |
| CPU Util. | 3.65 | 0.04 | 0.13 | 0.15 |

Table III: Server Metrics Area Over Threshold

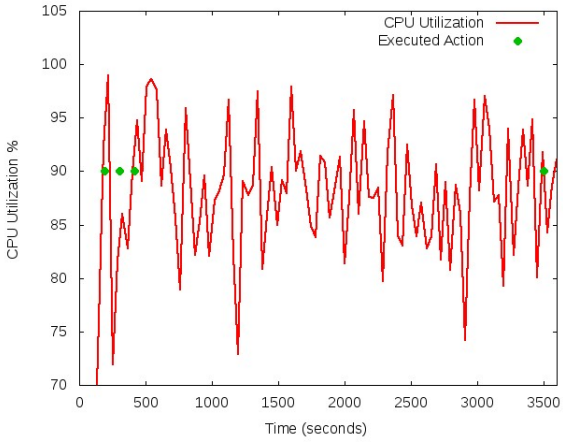



Figure 16: Example Run with Diagnosis - CPU Utilization

period when the server was operating under maximum load (75 clients total). The results are slightly different in value to those of the entire run, but values in comparison with each other remain consistent.

Another way to look at the data is to examine not the averages but the amount of time the value is over the specified threshold, and by how much. This can be done by calculating the area of the curve over the threshold. Figure 14 shows the measured response time of the Apache web server with the manager disabled and with weighted action selection, compared to the threshold value of 2000ms, for a single repition of the experiment. The area between the threshold value and the response time curve above it provides a useful measure of how well the goals of the policies are being achieved. Table III contains these values. The values shown are averaged over the 5 experiment repetitions. The system with the manager disabled exceeds the thresholds of both Apache response time and CPU utilization far more than with the manager enabled. Diagnosis favouring fewer actions comes out on top yet again, with weighted actions and diagnosis features multiple actions coming in second and third, respectively. Note that since the units of response time and CPU utilization are not the same, we cannot compare directly between the response time and CPU utilization area over threshold values.

## IX. Example Run with Diagnosis

To help illustrate how the autonomic manager behaves, particularly when using the diagnosis algorithm for action selection, we will take a look at an example experiment run and go into some detail at a few points of interest. The information examined is from a single experiment repetition, but is typical of all of the experiments. We will look at the diagnosis algorithm favouring hypotheses with fewer disorders (fewer actions to take). It is difficult to tell the direct consequences of each decision made and action executed, since a very large number of actions are executed throughout the experiment and the workload is dynamic. Nevertheless, some thoughts as to why the diagnosis algorithm performs slightly better than weighted action selection can be derived from such an analysis.

Figures 15 and 16 show the response time and CPU utilization metrics for an example run of the system using diagnosis favouring fewer actions. Four points of interest are marked on each graph and explained in some detail, in order from left to right (sequential order in time).

*Point 1:* The first violations occur at around the three minute mark (180 seconds).

1) Apache Response Time Violation
2) Apache CPU without Response Time Violation
3) PHP Response Time Violation
4) MySQL Response Time Violation

To begin with, this is an interesting combination of violation events. The first, third and fourth violations are all triggered by the same conditions, namely, the response time of the web server exceeding 2000ms and having an increasing trend. The difference lies in the set of advocated actions by each violation. Each policy advocates actions related to a different component of the system, namely, the Apache web server, the PHP cache, and the MySQL database. What makes this particular set of violations interesting is the second violation, namely, the Apache CPU without Response Time Violation. The conditions for this violation are CPU utilization above 90% and rising, and the response time of the web server being below 200ms, a contradiction with the conditions of the other policy violations. Clearly the response time cannot be both above 2000ms and below 200ms. What probably has occurred is that both states were present at some point in the interval between the last time the policies were checked for violations and this time. This interval during these experiments was 10 seconds.

The diagnosis algorithm then attempts to build hypotheses that can explain the situation we are seeing, even though we know that these particular violations do not represent a single snapshot of the state of the system, but rather what has occurred over the last 10 seconds. This is not necessarily a bad thing, as such seemingly contradictory information may in fact lead the diagnosis algorithm to finding a better solution by eliminating some extraneous actions or even including actions that may not have been considered otherwise. Whereas the weighted action selection will select an action based on applying some importance to each policy and each action independent of each other, the diagnosis algorithm takes into account the entire situation in its decision making. This may account for some of why the diagnosis algorithm performs better than weighted action selection.

The diagnosis algorithm builds the set of all possible actions or sets of actions that can cover the given policy violations. In this case, the first three are single actions that cover every condition in each policy. Since in this example the algorithm is favouring hypotheses with fewer disorders (fewer actions to take), these single actions are ranked first.

1) Decrease the maximum number of clients in Apache
2) Decrease the maximum number of KeepAlive requests in Apache
3) Decrease the maximum bandwidth, which compromises the performance of lesser service classes to maintain the performance of higher classes, with gold being the highest and bronze the lowest.

Hypotheses indicating that more than one action should be performed are ranked lower. An example of such a hypothesis is one that advocates both decreasing the MySQL Key Buffer size and increasing the cache memory available for PHP at the same time. The ordering of hypotheses containing the same number of actions is arbitrary, and is based on the order in which the violations are given to the algorithm and how the algorithm operates. It can be considered essentially random. Nevertheless, actions are attempted in the order they are sent to the PEP. In this particular case, the first action (decrease the maximum number of clients) was not performed because its associated test failed (the parameter was already at its lowest possible value). The second action, decreasing the maximum number of KeepAlive requests, was performed.

*Point 2:* After the first set of violations, a large number of the policy violation situations consist simply of the three Response Time Violations.

1) Apache Response Time Violation
2) PHP Response Time Violation
3) MySQL Response Time Violation

Since all three of these violations share the same conditions, the resulting diagnosis is simply a list of all actions advocated by the three policies, because any of these actions will cover all of the conditions of all three. Since the diagnosis algorithm performs no ordering of the actions within itself, it will build the same set of potential actions as the weighted action selection method, except it will make no attempt to determine which is more likely. As such, it will probably make a similar, if not slightly worse decision. The tests attached to the actions also make a difference in which action is selected, as all tests for an action must pass before the action can be executed. This means that several higher ranked actions may be skipped before reaching an action that can be performed, potentially neutralizing some of the effect of ordering.

*Point 3:* Another common policy violation situation occurs at the 417 second mark. At this point, we see a combination of both response time related violations and CPU utilization violations.

1) Apache Response Time Violation
2) PHP Response Time Violation
3) MySQL Response Time Violation
4) Apache CPU Utilization Violation
5) PHP CPU Utilization Violation
6) MySQL CPU Utilization Violation
7) Apache CPU and Response Time Violation

We have already seen the response time violations. All three contain the same conditions but advocate actions related to different components of the system. The three CPU Utilization violations (4, 5 and 6) are similarly related. All three have CPU utilization above 90% and an upward CPU utilization trend as their conditions, but they each advocate different actions. The Apache CPU and Response Time policy violation is triggered by a combination of both web server response time conditions and CPU utilization conditions, and advocates actions to be taken in the case that both the response time is above 2 seconds and CPU

utilization is above 90%. This policy attempts to dictate what should occur when more than one type of violation exists, and the set of actions advocated by it is actually a subset of the actions already advocated by the other policies. Such a policy is essentially trying to simulate some sort of diagnosis, and is likely rendered obsolete by the diagnosis algorithm. Nevertheless, it is in use at the moment and taken into consideration in diagnosis. The following is the list of actions or sets of actions returned by the diagnosis algorithm.

1) Decrease the maximum number of KeepAlive requests in Apache
2) Increase the cache size used for PHP pages
3) Decrease the maximum bandwidth
4) Increase the MySQL thread cache size and increase the number of Apache clients
5) Decrease the maximum number of clients in Apache and increase the MySQL key buffer size
6) Increase the MySQL thread cache size and key buffer size
7) Decrease the maximum number of clients in Apache and increase the MySQL query cache size
8) Increase the MySQL query cache size and thread cache size

As before, hypotheses containing fewer actions to perform are preferred. Only one of these will be executed, and they will be attempted in the order listed. Again, the ordering within hypotheses containing the same number of actions is essentially random.

*Point 4:* At around the 59 minute mark another interesting policy violation situation occurs.

1) Apache CPU Utilization Violation
2) PHP CPU Utilization Violation
3) MySQL CPU Utilization Violation
4) Apache without both CPU and Response Time Violation

We have already seen the three CPU Utilization policy violations. The fourth, Apache without both CPU and Response Time, indicates that response time is within normal constraints (below 2 seconds), and that CPU utilization is also below the violation threshold of 90%. Clearly, as we saw earlier with response times, this contradicts the other three policy violations, again most likely due to the 10 second window in which violations can occur before they are processed. The question then becomes, how should this be interpreted? This is by no means a trivial question. Should the violations indicating that the CPU utilization is over 90% be trusted or the one indicating that it is below be trusted? In weighted action selection, one of these two options will be chosen. With diagnosis, however, both options will be combined to find some solution that satisfies both, thus taking into account all of the information received. Such a difference in approach may be at least partially responsible for the improved performance of the diagnosis algorithm. In this case, a set of four hypotheses is generated, each containing two actions to perform.

1) Decrease the maximum number of clients in Apache and increase the maximum bandwidth
2) Decrease the maximum number of KeepAlive requests in Apache and increase the maximum bandwidth
3) Increase the cache size used for PHP pages and increase the maximum bandwidth
4) Increase the MySQL thread cache size and increase the maximum bandwidth

As before, the actions were attempted by the PEP in the order shown, and in this case, the very first set of actions passed its tests and was performed.

## X. Conclusions and Future Work

A diagnosis approach using adbuction has been proposed to help the autonomic manager decide which action to take in the case of multiple policy violations. The approach uses the policies themselves to build a Causal Network, which is then used to perform diagnosis. The diagnosis algorithm was implemented in the BEAT Autonomic Manager [3] for testing.

We examined the performance of a web server without the aid of the autonomic manager, with the manager using weighted action selection, and using diagnosis. From the results presented here, we can conclude that the diagnosis algorithm performs at least as well as the previous method of action selection (weighted action selection). CPU utilization for all three action selection methods stays below the threshold, but the two diagnosis methods make use of more CPU resources than weighted action selection, keeping closer to the threshold. Diagnosis favouring multiple actions performs similarly to weighted action selection, except on the actual measured client response times, where it has an edge on gold and silver service class response times. Diagnosis favouring fewer actions beats out the other methods across the board, although not by a significant margin. This indicates that the use of the diagnosis algorithm to select an action in the case of multiple policy violations makes better decisions than the previously developed weighted action selection methods, more closely achieving the overall goals of the policies, that is, keeping metrics such as CPU and Response Time within specified thresholds. Although the improvement offered by diagnosis was minor, in a larger scale experiment it may become more pronounced. Further experimentation is required to fully evaluate the method.

The advantage that diagnosis favouring fewer actions has over diagnosis favouring multiple actions seems to indicate that simpler explanations of the given set of policy violations (hypotheses containing fewer disorders) are more likely to be correct, an example of Occam's Razor [19]. The decision making advantage enjoyed by diagnosis over weighted action selection may be due to the fact that diagnosis essentially attempts to use all available information together

to make a decision, while weighted action selection pits each option against each other. This is a subtle yet potentially interesting distinction.

The diagnosis method is not a strict alternative to weighted action selection, and future work could investigate the combination of these methods. The criteria for policy and action weighting could be used to build probabilities into the causal network. The policies themselves should also be examined, as a simpler set of policies may be possible when using diagnosis. Finally, in order to fully evaluate and drive development of these techniques forward, some larger scale implementation and testing is likely necessary.

## REFERENCES

[1] J. Kephart, D. Chess, I. Center, and N. Hawthorne, "The Vision of Autonomic Computing," *Computer*, vol. 36, no. 1, pp. 41–50, 2003.

[2] L. Lymberopoulos, E. Lupu, and M. Sloman, "An Adaptive Policy-based Framework for Network Services Management," *Journal of Network and Systems Management*, vol. 11, no. 3, pp. 277–303, 2003.

[3] R. Bahati, M. Bauer, E. Vieira, O. Baek, and C. Ahn, "Using Policies to Drive Autonomic Management," in *WoWMoM 2006. International Symposium on a World of Wireless, Mobile and Multimedia Networks.*, 2006, pp. 475–479.

[4] R. Bahati, M. Bauer, and E. Vieira, "Policy-driven Autonomic Management of Multi-component Systems," in *Proceedings of the 2007 conference of the center for advanced studies on Collaborative research*. ACM New York, NY, USA, 2007, pp. 137–151.

[5] M. Tighe and M. Bauer, "Mapping Policies to a Causal Network for Diagnosis," *ICAS 2010, International Conference on Autonomic and Autonomous Systems*, pp. 13–19, 2010.

[6] Y. Peng and J. Reggia, *Abductive Inference Models for Diagnostic Problem-solving*. Springer, 1990.

[7] N. Damianou, N. Dulay, E. Lupu, and M. Sloman, "Ponder: A Language for Specifying Security and Management Policies for Distributed Systems: The Language Specification," *Imperial College Research Report DoC*, 2000.

[8] R. Anthony, "Policy-centric Integration and Dynamic Composition of Autonomic Computing Techniques," in *Autonomic Computing, 2007. ICAC'07. Fourth International Conference on*, 2007.

[9] J. Lobo, R. Bhatia, and S. Naqvi, "A policy description language," in *AAAI '99/IAAI '99: Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence*. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1999, pp. 291–298.

[10] Distributed Management Task Force, "Common Information Model Simplified Policy Language," http://www.dmtf.org/standards/cim_spl/, June 2010.

[11] A. Bandara, E. Lupu, and A. Russo, "Using event calculus to formalise policy specification and analysis," in *Policies for Distributed Systems and Networks, 2003. Proceedings. POLICY 2003. IEEE 4th International Workshop on*, 2003, pp. 26 – 39.

[12] S. Duan and S. Babu, "Guided problem diagnosis through active learning," in *Intl. Conf. on Autonomic Computing*, 2008, pp. 45–54.

[13] S. Ghanbari and C. Amza, "Semantic-driven model composition for accurate anomaly diagnosis," in *Autonomic Computing, 2008. ICAC'08. International Conference on*, 2008, pp. 35–44.

[14] J. Bigus, D. Schlosnagle, J. Pilgrim, W. III, and Y. Diao, "ABLE: A Toolkit for Building Multiagent Autonomic Systems," *IBM Systems Journal*, vol. 41, no. 3, pp. 350–371, 2002.

[15] R. Bahati, M. Bauer, and E. Vieira, "Adaptation Strategies in Policy-Driven Autonomic Management," in *ICAS 2007, International Conference on Autonomic and Autonomous Systems*. IEEE Computer Society Press, Washington, DC, USA, 2007, p. 16.

[16] Y. Peng and J. Reggia, "Plausibility of Diagnostic Hypotheses: The Nature of Simplicity," in *Proceedings of AAAI-86*, 1986, pp. 140–145.

[17] "PHP Bulletin Board," http://www.phpbb.com/, January 2011.

[18] "Apache JMeter," http://jakarta.apache.org/jmeter/, January 2011.

[19] "Merriam-Webster Online Dictionary, Occam's Razor entry," http://www.merriam-webster.com/dictionary/Occam's razor, May 2009.

# A Framework for Progressive Trusting Services

Oana Dini
University of Besançon, France
oana.dini@univ-fcomte.fr

Pascal Lorenz
University of Haute Alsace, France
lorenz@ieee.org

Hervé Guyennet
University of Besançon, France
guyennet@lifc.univ-fcomte.fr

***Abstract* – The web-based transactions, web services, and service-oriented platforms require appropriate mechanisms to announce, select, and use different services. A user is always under the dilemma of 'use and trust' or 'trust and use' for different services based on the notion of service reputation. The interaction between every service provider and its users is regulated by the service level agreement and customer satisfaction feedback. The former is the basis for the technical audit, while the latter subjectively validates the user perception. Selection of a most appropriate service by correctly invoking is a challenge. This is due to the difficulties to correctly expose proper way to invoke a service, to the variety of services, from on-line services, software pieces, to shopping, and to different invoker behavior. When considering invoker feedback, service ranking based on user perception, or based on recommenders' statistics are relevant. A significant aspect is played by service similarity. The paper presents a framework and appropriate mechanisms to evaluate the services/providers in the light of their respective direct impact on user perception. To accurately evaluate the feedback after service/product consumption, we will refine the user profile by considering the dynamics of the feedback. The approach we propose deals with peaks in feedbacks. We consider quick negative and quick positive feedback as well as late vs. early feedback with respect to the time of the transaction. We propose formal concepts used in selecting an appropriate service. The paper presents adapted approaches to select services based on distance and similarity, and introduces a similarity taxonomy to better tune various kinds of service invocation under specific constraints, such feature relaxation, type of similarity, context, and service ranking. Selection is based on the feedback from the user. The proposed model is used for building a selection algorithm that allows variations on service invocation. The proposal is validated through use cases.**

***Keywords – recommenders; reputation; dynamic feedback; services similarity; temporal similarity; use profiles, dynamic patterns.***

## I. INTRODUCTION

With the overwhelming amount of information, products, and services available over the Internet, it has become harder for the users to select the ones that fit best their needs or requirements. First of all, it is too difficult and time consuming to sort through hundreds of items and select the needed one. Also, there is the problem of trusting the provider for that item and not only that, but trusting that the provider is offering a product that meets the user's requirements. In order to assist the user in selecting the product or service that it needs, recommender systems have been proposed.

Recommender systems (RS) have been the subject of many studies and products over the last decade. The term was first brought up by Resnick and Varian [1], which, as mentioned in [2], it was mostly a replacement for "collaborative filtering" proposed in [3]. Recommender systems are defined as systems that collect ratings from users and then analyze the data to produce recommendations to other users [4]. There are several techniques used to generate recommendations, but the main categories are Content-based Filtering (CBF), Collaborative Filtering (CF), and Hybrid approaches [5].

RS are important in electronic commerce, especially for marketing [6] and they have been widely used in order to attract and retain customers. The relation between the loyalty of users and RS was studied in [7] using data from Amazon.com. Their findings showed that the presence of consumer reviews helps with retaining customers and also attracting new ones. In time, the business gains reputation, which usually translates to increase in business. There are a few challenges in optimally using the recommenders due to the variety of user's profile and its volatility and the reputation of different service providers. For dealing with these aspects, recommenders usually use product rating, confidence in service providers, and regularly update this information for an accurate suggestion for a given request.

In all the existing approaches, some improvable assumptions are considered for the purpose of easily

computing the reputation. Aspects like partial feedback, ignorance of customer confidence, and most importantly, lack of information on the service provider identity are major challenges for an accurate reputation per product, per service provider, per context, or per user profile. In this paper, we propose an approach taking into consideration the above challenges and deriving mechanisms for a more accurate reputation.

Classically, two notions concerning the quality of a delivered service are correlated for an accurate service evaluation, i.e., QoS (Quality of Service), and QoE (Quality of Experience). Specific to each service, there are particular service parameters that are agreed upon between a provider and a subscriber, commonly settled by the SLA (Service Level Agreement). On the provider side, the SLA parameters are used for technical audit and litigations (leading to penalties or bonuses towards a given user or class of users). Specific on-line and off-line measuring mechanisms for SLA metrics and specialized audit techniques have been proposed. On the consumer side, the subscribers' satisfaction is gathered and mapped to the audit results to validate a given service, to detect flaws in delivering a service, and to ultimately build a view on service reputation. In general, a record is handled per service or per products, with respect to a given subscriber or a class of subscribers. Customer feedback can be 'by request', or 'at will', and embraces various forms of on-line questionnaires. As a result, a customer might decide not to answer, or to answer exhibiting a particular behavior. Ultimately, a service provider might fake some feedbacks to increase its reputation. There are various factors that influence the computation of an accurate reputation, e.g., the volume of ordered services, the diversity of the subscriber classes, customer trust and loyalty, and the dynamics of the feedbacks. Practically, the main problem we try to find a solution for is to dynamically and accurately compute the reputation of a service/product, based on the system transactions. We propose a simple formula for reputation updates (1). The challenge is to identify the correct metrics in computing the updated reputation.

$$r_{real} = (1+ \lambda) \times r_{current} \quad (1)$$

where:
$r_{real}$ represents the updated reputation, considering
$r_{current}$ represents the known and accepted reputation, and
$\lambda$ represents the correction based on customer perception and feedback behavior, $\lambda$ belongs to $\{(-\alpha, \alpha) \mid \alpha > 0\}$, usually having values in the vicinity of '0'.

The main achievement of our proposal is the following. The recommenders systems have a powerful mechanism to accurately indicate the real reputation, when selecting the best service provider from a service directory. Except some studies on QoE [17][19] that mainly consider technical metrics, we introduce and evaluate the customer behavior.

The paper focuses on dynamic aspects of customer feedbacks and formalizes mechanisms for a more accurate evaluation the reputation of a given service delivered by a given provider in establishing policies for $\lambda$.

The large spectrum of user behaviors (and, in general, the variety of needed services) leads to the need of similarity-based matching, when a given service is required. Traditionally, the notions QoS and QoE deals with these aspects. However, the perfect matching and the approximate-matching depend on a large number of factors. For example, if we consider Web Services dedicated to weather forecast, location, month/day/year, parameters (rain, wind, temperature, and pressure) can be appropriate parameters when inquiring. Definitively, there are several forecast services, and the experience of a particular user might differ from one forecast service to another. Some provide information that is more accurate than others (i.e., data is more frequently updated), history is better preserved by particular services, via backward search, e.g., Weather Underground, etc. A similar problem is observed when choosing and downloading a particular piece of software, when inquiring for a specialized on-line book shop, or for looking for a service providing the most updated world-wide information. Finally, some services offer a friendlier interface to search, order, and get delivered a particular need (i.e., personalized interface, myAccount, etc.).

There are meta-services, providing the service at choice. Such examples are those for buying flight tickets, where the cheapest, the quicker, or other selection criteria are used for service selection. Other meta-services are for selecting the most appropriate software to download, or to book a hotel. In most of the cases mentioned above, one criterion is usually considered to select from an existing service pool.

Two phases involve service features, (i) service discovery (locating) and (ii) service selection (in the case of a set of services, relatively satisfying the needs with similar degrees of satisfaction). Both phases require special mechanisms to assess service similarity. Meta-services have a restraint number of known services, well localized, the parameters of which are also in small number. Then, selection appears to be less complex. With a well known service and limited (usually one search/ selection criterion) similarity is relatively easy to be determined. The above considerations are not longer

valid for a large spectrum of properties a service might expose to satisfy a given service request. To satisfy a request, service similarity plays an important role for timely identifying and delivering, and for an optimal (maximal) customer (invoker) satisfaction. Customer satisfaction is expressed by QoE, on-line feedback, service ranking, and manifested by variations of QoS to keep service costs and satisfaction in synchrony.

The paper is structured as follows. Section II covers related works. An enhanced recommender model is presented in Section III. In Section IV presents a taxonomy of the dynamic feedback and a dual architecture. Section V introduces computation mechanisms for an accurate reputation of a service via dynamic reputation update policies and heuristics. A context-based similarity model, including distance and similarity metrics, a similarity taxonomy, and other facilities to consider service ranking and feature relaxations is introduced in Section VI. Section VII presents an algorithm to compute a minimum set of existing services satisfying a given query, following the newly introduced model. Section VIII concludes the paper and presents future investigations.

## II. RELATED WORKS

As the proposed approach touches the recommendation and reputation on recommenders, service providers, and products, we first introduce some basic concepts.

### *2.1 Concepts*

The core information of a recommender is a list of offers (products) and ratings of those products based on feedback received after a series of recommendations. The *rating* is subject to incomplete, fictitious feedback, volume of transactions for a given product or provider, and confidence in feedback. Based on the ratings, the recommender computes its own *ranking* per product.

*s[r], P[r]* represents a service or a provider with the rank r, where r is an integer.

Associated with the ranking is the notion of *reputation* that in fact determines the ranking. The reputation formula, while product oriented, it might not be accurate, as its computation cannot avoid some realities, such as some service providers have private relationships with recommenders (e.g., publicity, sponsorship) or indirect servicing (recommended product might not be produced by the front end provider, but simply delivered by it).

Reputation is an index associated with the service or a product based on user feedback that is taken into consideration when the ranking is calculated. The reputation index usually belongs to a set, *{outstanding, very good, good, acceptable, bad}***.** A recommender might increase the rank of a service when its reputation index, for example, passes from very good to outstanding [8]

*Similarity* is another concept used in generating recommendations. In order for a recommender to suggest products to a user, it needs to find a commonality among users (this applies in the collaborative approach) or among the products that were rated in the past by the user (this applies in contend-based approach). There are different techniques used to compute the similarity measure, but the most used are correlation-based and cosine-based techniques [5] [9]. Similarity is an index associated with two services or products. For example, $\mathbf{s_1}$ **[~/80%]** $\mathbf{s_2}$ means $s_1$ is similar with $s_2$ with an acceptance of 80% based on the service's features or in the same range of ranking.

### *2.2 Current approaches for recommenders*

Recommenders are usually classified based on the approach for making the recommendations. There are three main categories of recommender types: content-based filtering, collaborative filtering, and hybrid filtering.

The *content-based filtering* recommends to users items that are similar to the ones searched by the users in the past [5][9]. This type of recommendation technique is mostly used to recommend text-based items such as documents and newspapers. In order to produce the recommendations, the system needs a profile of the user, which is represented by a set of terms. The profile can be obtained from the user through a questionnaire or it can be learned from their past transactions. This type of filtering has its shortcomings. Since it is content-based, it needs to have the representation of data in a matter that can be machine-parsable (e.g., article). It is harder to apply this technique in the case of movies, music, images, which are not machine-parsable.

The *collaborative filtering* (CBF) [3] tries to predict the relevance of an item based on the ratings done by other users. It accumulates ratings of products and whenever a request comes, the system identifies similar users and recommends the products rated by them. In this type of filtering, the user profile is defined by a vector of items and their ratings, which is updated over time. As opposed to the CBF, this type of filtering can

be applied to any kinds of items, not only to machine-parsable items. However, there are limitations with this approach, mostly caused by the lack of data points in initial stages: new user and new item.

The *hybrid algorithms* usually combine the content-based and the collaborative algorithms to overcome some of the limitations of the other two approaches. This approach has been adopted by some RS [10], [11]. There are different ways to combine the two algorithms and [5] present the different approaches in detail.

As we mentioned above, the reputation of a business is gained in time, mainly based on reviews from users. This brings up another point and that is obtaining accurate reviews from users. Many users are not willing to leave feedback after a transaction is completed. One reason for not leaving feedback is the lack of incentives. If there isn't some kind of payoff for the feedback, the user won't put the effort into posting one. An incentive mechanism is addressed in [12] where incentives are given to users who provide honest feedback through a side payment mechanism. Examples of incentives mechanisms are Amazon's "Top Reviewers" practice and Epinions.com referral fees practice [13]. Another reason for not leaving feedback is to purposely withhold information about a product that gives its user an advantage [14].

Another concern related to the validity of the reviews is the manipulation of the reviews by parties with direct vested interests. Businesses can review their own products in order to boost the sales. Also, the competition can leave or fabricate negative feedbacks to undermine the competitor's reputation. There are ways to filter out biased feedbacks and to prevent manipulation [15], but preventing coordinated collusion attacks is still an issue. eBay, for example, does not have a problem with feedback manipulation. The feedbacks can only be left by users who are registered with them and who made a purchase on eBay. However, if a group of users agree with a seller to leave positive feedback for fictitious auctions (e.g., the seller can post multiple 1 cent auctions on which the users can bid), the seller's ratings can be positively affected. These users are usually called shills. This approach would require quite an effort (the larger the number of shills, the bigger the impact), but it can be achieved.

Reputation is very useful in RS and eBay is one example of a reputation system that proves that their approach works well. However, having a centralized reputation system, such as eBay, can bring other issues, such as vulnerability and inflexibility of the system [14]. In [14], the authors propose a distributed trust and reputation management framework. The users choose a trust broker and after each transaction with a service, the user sends its rating to its trust broker. This way, the trust broker builds a reputation about a service based on the user's feedback. The brokers exchange reputation information among themselves in order to collect more information about the available services. This framework relies on the user's feedback only, ignoring the business model of the provider.

Recommender mechanisms [18] rank the products or services based on feedback received after a series of recommendations and successful transactions. The *rating* is subject to incomplete, fictitious feedback, volume of transactions for a given product or provider, and confidence in feedback. Based on the ratings, the recommender computes its own *ranking* per product, defining the reputation (r) of a service/product. A computational mechanism including user' confidence (c) and feedback expectation (e) was proposed in [16]. An attempt, rather static, of considering a static subjective evaluation of the quality of the voice service is described by the MOS (Mean Opinion Score). The MOS is an arithmetic value ranging between 1 and 5, expressing individual perception [17]. However, MOS apply strictly to voice-related services, on an individual basis. The metrics are purely technical and related to codec use, packet loss, packet reorder, packet errors, and jitter. Another standard for evaluating the speech QoE, also considering technical metrics, is captured by PESQ algorithms [19].

A dynamic approach for customer input is presented by byClick system [20][21], where there is an on-line click-counting on the number of service accesses. No customer behavior, subscription status, or feedback patterns are considered. However, in the current approaches, no correlation with the frequency of users' report and transaction peaks, as well as with the users' report patterns were considered.

Finding similar services (approximate, but satisfactory matching) is somehow similar to (i) text matching, (ii) schema matching, or (iii) software-component marching. For some text matching solutions (information retrieval) mechanisms based on term frequency are used [28][29]. In schema matching, special techniques are using semantics of the schemas to suggest schema matching [30]. Mainly, linguistic and structural analyses, as well as domain knowledge, are methods to handle schema matching. When expanding to software component matching [31] (considerably used in software reuse) component signature and program behavior (usually formally defined) are considered; in this case, data types and post-conditions should be considered for matching. However, these techniques are not suitable for Web

Services [27], as data types and post-conditions are not available. Usually, such a service has a name and text description in UDDI **(**Universal Description, Discovery, and Integration) registry, operation decryptions, and input/output descriptions; the last two are usually specified in WSDL (Web Service Description Language).

Dong *et al.* [27] proposed criteria for associating similar terms. They introduced the cohesion/correlation score, as a measure of how thing two terms are. However, they do not consider particular characteristics of a term. They applied the score only to Web Services. We start from the point that services similarity has a meaning only between services than can be context-oriented and belong to a cluster (e.g., invoking a service gives a list of similar operations with similar results). Other approaches consider both diversity and similar at the same time having the distance as a metric [32]. We adopt these metric (see Section III) and adapt them to the service similarity computation.

In fact, specific to each service, there are particular service parameters that are agreed upon between a provider and a subscriber, commonly settled by the SLA (Service Level Agreement). On the provider side, the SLA parameters are used for technical audit and litigations (leading to penalties or bonuses towards a given user or class of users). Specific on-line and off-line measuring mechanisms for SLA metrics and specialized audit techniques have been proposed. On the consumer side, the subscribers' satisfaction is gathered and mapped to the audit results to validate a given service, to detect flaws in delivering a service, and to ultimately build a view on service reputation. In general, a record is handled per service or per products, with respect to a given subscriber or a class of subscribers. Feedback can be used to enforce service similarity.

In this paper, we also expand the cluster-based similarity to service similarity and introduce similarity taxonomy, where the service consumer has a weight in deciding service similarity. The idea is to establish service ranking (and reputation) inside a given cluster, and define similarity considering service-provider and service-user feedback.

## III. AN ENHANCED RECCOMENDER MODEL

In this section, we present a recommender model that can handle the sub-contract mechanism, yet keeping an accurate information on a given provider reputation (leading to an accurate ranking).

### *3.1 Setting the case*

A simple scenario is presented in Figure 1, where the user is interested in service $s_1$ from $P_1$. The user asks the Recommender for the best provider for service s1 within specific parameters. The Recommender replies with either a provider that has the best reputation for service $s_1$ or with a list of providers $\{P_i\}$ for $s_1$. Let us assume $P_1$ is registered of being capable to deliver $s_1$ (others might be registered for $s_1$ as well). The Recommender cannot know if $P_1$ has the service or if it contracts it from a different provider. If $P_1$ is contracting $s_1$ from $P_2$, the transaction between $P_1$ and $P_2$ is transparent to both the Recommender and the user. At the end of the transaction, the user sends the rating of $P_1$ to the Recommender and $P_1$ receives all the credit for the transaction. This leads to an inaccurate reputation and altered ranking.

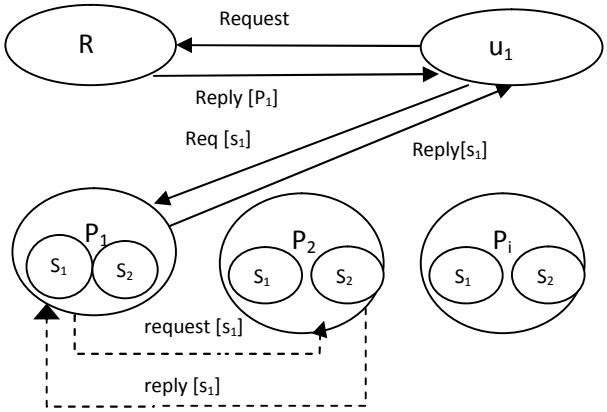



Figure 1. Indirect reputation

If the reputation of the provider is based only on the user's feedback, there is no way to assess the ultimate role of each provider. In order to have a more accurate picture of the providers' involvement, we propose that feedback from the providers be taken into account when establishing reputation. This includes both the front end provider (in our case $P_1$), as well as any subcontracted providers (in our case $P_2$). All feedback goes directly to the Recommender.

The ideal scenario would be when all the users and providers report 100% of the transactions. In reality, users don't always leave feedback and providers do not always report rendered services. In such a case, the Recommender is left to deal with an incomplete set of data. Moreover, some of the reported data may be fabricated by both users and providers.

*3.2 Recommender representation model*

Apart from the mechanism of collecting the feedback and interfacing with the users, the core information present in a recommender is stored in a service database. This allows a request to be replied to with a service or a list of services, eventually with a degree of similarity associated with each service. Usually, the recommender keeps information on relative ranking among these entities.

We propose an enhanced model, which takes into account the user's profile and behavior, and a list of potential providers for a given service. This allows a more refined ranking scheme where providers can be rated per service.

While ranking is based on user feedback, there is no appropriate mechanism to consider the user's expectation (e) and credibility (c). By user expectation we mean the probability of having the user leave feedback after a service was delivered. The credibility refers to the user's ability to give a trusted rating. Usually both, expectation and credibility are expressed as percentage.

In Figure 2, we present the enhanced recommender model. The recommender stores information about the available services, the providers and their services, plus the user profile, which includes its expectancy and credibility. Both services and providers are associated with a rating. The providers' rating is done within the context of a service. This way, the rating can be done per product and per provider for a specific product.

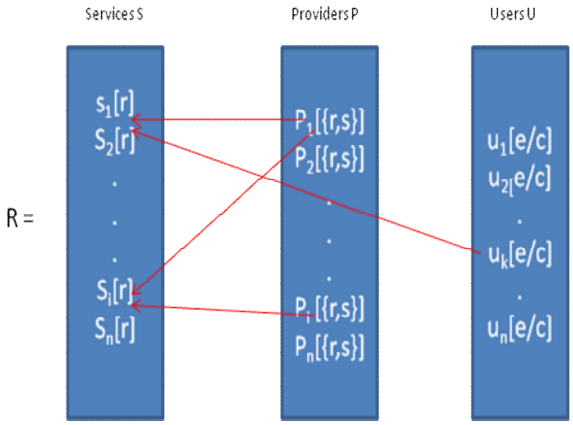



Figure 2. Enhanced Recommender Model

By keeping the relationships between the providers, their services, and also the users who requested the available services, the recommender can provide better suggestions and answer to more complex queries.

We classify queries in two categories, i.e., U-R and P-R. Some salient queries U-R might be:

Query 1:

input: $[s_1]$
output: $[s_1/P_1, s_1/P_2]$

The user asks for service $s_1$ and the recommender replies with a list of providers that offer $s_1$.

Query 2:

input: $[s_1]$ & $[s_1 (\sim/\mathcal{E})]$
output: $[s_1/P_1, s_1/P_2]$ & $[s_i/P_i]$

The user asks for service $s_1$ and/or a service similar to $s_1$. The recommender replies with a list of providers that offer $s_1$ and/or a list of providers who offer services similar with $s_1$.

"$\sim/\mathcal{E}$" represents the similarity of services with $\mathcal{E}$ as proximity

Query 3:

input: $[s]$ $[P_1, P_2]$
output: $[s_1/P_1, s_2/P_1]$ $[s_i/P_2, s_j/P_2]$

The user asks for a list of services offered by certain providers. The recommender replies with a list of services offered by those providers.

Query 4:

input: $[ s \mid r > x]$
output: $[s_1/r_1, s_2/r_2]$

The user asks the recommender for a list of services, which has a ranking "r" higher than a certain value. The recommender replies with the list of services.

Some relevant queries P-R might be the following:

Query 5:

input: [$u_i$]
output: [$u_i$ [e/c]]

The provider asks the recommender about user $u_i$. This may be relevant to the provider in order to assess the user's credibility. The recommender replies with the $u_i$ expectation "e" and credibility "c".

Query 6:

input: [ all $U_i$, e > α, c >β]
output: [$u_i$ [e/c]]

The provider asks the recommender for a list of users whose expectation and credibility are higher than a certain value. This may be relevant to the provider in order to assess the user's credibility. The recommender replies with the list of user(s).

Based on the formula presented in the following section, complex information can be gathered and more accurate answers to different queries can be provided.

*3.3 Computation mechanism*

The enhanced model allows a more comprehensive schema for computing the reputation.

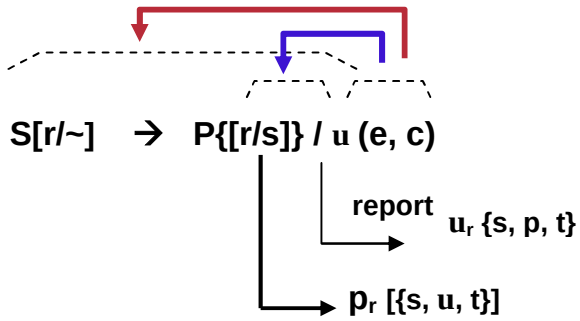



Figure 3. A computation schema for recommenders

In our framework, a recommender has mechanisms for representing services (S) with their reputation (r) and similarities (~), provider (P), with their reputation (r) linked to the reputation of their service providers (s), associated with user's (u) expectation (e) and credibility (c). A particular relation is valid at a moment (t). For example, a user x is expected to provide feedback with e = 80% and the confidence on its feedback is 70%. The feedback is on a provider (p) providing a service (s) at the time (t). The schema allows having a reputation view of a user at a given time, on a given provider delivering a given service. The schema also allows having a reputation of a provider, as perceived by a user at a given time, if delivered by a given service.

We are now going to concentrate on different scenarios dictated by the amount of data reported by users and service providers.

For example, a user sends a request to the recommender for the best cell phone provider that would meet certain parameters. The recommender replies with provider $P_1$. The user makes a request for a number of cell phones from $P_1$. After the transaction is completed, all the involved parties have the option to send feedback to the recommender. The recommender collects the data and based on the feedback, it updates the reputation of the involved parties. The nature of the collected data can be divided in three main cases:

*3.3.1 Matching reports*

The number of feedback reports from the user matches the number of reports from the service provider within a particular time window relevant to the service type. To continue with the example from above, the user sends the feedback to the Recommender, including the number of cell phones that it purchased. $P_1$ reports to the Recommender that the user purchased a number of cell phones from it. The numbers reported by both the user and $P_1$ match.

A subclass of this scenario would be when $P_1$ sub-contracts from a different provider, $P_2$. If $P_1$ receives a request for cell phones, it can send the products from its own stock, send part from its own stock and part from $P_2$, or get the entire order from $P_2$. In this case, the Recommender would receive reports from both providers, $P_1$ and $P_2$. The exact number reported would not match since $P_1$ will report that it sent the entire order to the user, and $P_2$ would report that it sent a certain number of phones to $P_1$, but the data can be correlated. The correlation is done by using the transaction completion time, the user identifier, and the provider identifier.

*3.3.2 Over-reporting providers*

The number of feedback reports from user and provider does not match. This can be caused by either providers exaggerating the amount of transactions completed, or by users who underreport. In this case,

some of the data can be correlated by the Recommender.

*3.3.3 Underreporting provider*

The number of feedback reports from user and provider does not match. This can be caused by either providers that do not report every transaction, or by users who exaggerate the amount of transactions completed. In this case, the Recommender can correlate some of the date.

*3.3.4 Case study for reputation correction*

Let us consider the following situation:

u -> [t] [p1] [s1], where u is the user, p1 and p2 are providers, t is the time of the request, and s1 is the service;
p1-> [t] [u] [s1], with p1 [r1/s1]
p2 → [t][s1], with p2 [r2/s1]

and the following transaction reports:

|u|: reports $\alpha$ transactions
|p1| reports $\beta$ transactions (with $\beta < \alpha$)
|p2| reports $\gamma$ transactions

then

$k = (\beta - \gamma) / \alpha$

In this case, for a given user u, and for the considered service s1, the real reputation is $r_1' = k \times r_1$, as there is an indirect service delivery form p2 via p1 to the user u. The schema allows having a more accurate view on who is delivering a service. Note that the number of transactions can be either reported or obtained by audit. In this use case, we consider that the providers are subscribed to an automated transaction report when delivering a service.

*3.3.5 Discussion*

In this section, we are comparing existing recommender systems with our proposal, on the basis of three main features: expectation, credibility, and user profile, as defined in Section 3.2.

We consider a few well known recommender systems and only selected those three main features as a basis of comparison. The existing recommenders do not incorporate in the user profile the expectation and credibility of a given user.

Table 1. Feature based comparison of several recommender systems as well as the proposed one

| | eBay | Amazon.com | Barnes & Nobles | proposal |
|---|---|---|---|---|
| expectation | Not in profile | Not in profile | Not in profile | Included in profile |
| credibility | Not in profile | Not in profile | Not in profile | Included in profile |
| User profile | yes | yes | yes | yes |

While the considered systems (eBay, Amazon.com, Barnes & Nobles) make use of the notion of profile when recommending a product, the main target is to identify potential similar services and products to either satisfy a request or recommend a particular service unknown to the user (using the similarity concept).

By including these features, the recommender can have a more complete view on user's satisfaction based on more accurate information maintained by the system on the user's behavior (the degree of responsiveness of the user ability to give trusted rating).

The performance and accuracy of a recommender system can be enhanced by including in the user's profile the user's expectancy and credibility. By having the expectancy of a user to leave a review and also its credibility, a recommender can better tune its suggestions to a user's requests with increased certainty. Ongoing experiments will identify the thresholds from where these features increase the accuracy of recommendations. Particular consideration will be given to the dynamics of user's feedback in terms of relationships between the frequency (volume) of the used services or products and the accuracy of the timely feedback.

## IV. DYNAMIC FEEDBACK

The mapping QoS~QoE principally involves SLA' metrics. In our approach, we also introduce temporal and ethical metrics to quantify more accurately the customer feedback. Additionally, long term and short terms feedback patterns are identified, including spikes feedback.

We consider the basic introduced in [16], where by user expectation we mean the probability of having the user leave feedback after a service was delivered. The credibility refers to the user's ability to give a trusted rating. Usually both, expectation and credibility are expressed as percentage. The new mechanism proposed includes the status of the user and the feedback history.

Finally, we drive feedback-based polices for service reputation updates, by considering these metrics, based on extreme behaviors of customers in terms of feedback, e. g., feedback too late, too quick, too frequent, too rare, etc.

### A. Dual reputation update architecture

Two views on reputation must be correlated for a given service/product, i.e., the provider view and the customer view. On the provider view, the perception of the service reputation, $r_{expected}$, represents the variation of several metrics, as the volume of sales, the number of new customers or lost customers in a given period. From the customer side, $r_{feedback}$ gathers customer perception on the reputation of a service.

Therefore, we propose a dual architecture to correlate and synchronize the two views. From the uniformity reasons, the update heuristics will follow a similar computation approach for both views, implemented by specialized engines.

Figure 4 depicts the main architecture and decisional and computation engines.

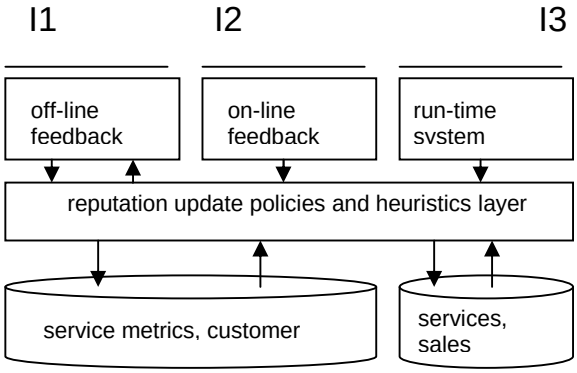



Figure 4. Dual reputation update architecture

The architecture considers two customer-facing interfaces (I1 and I2) handling the off-line (by_request, or at_will), and on-line (at_will), respectively, reputation updates. I3 is considering the provider-facing reputation updates. Three appropriate *off-line*, *on-line*, and *run-time* engines deal with the updates, by receiving and computing them via I1, I2, and I3, respectively. In the current paper, we only focus on aspects related to data collected via I1 and I3. For data collected via I3, an interesting implementation, nor related to our model, is presented by byClick [20][21]. While dynamic is different, the dual architecture can also support this approach. Hereafter, we will only refer to the off-line and run-time engines.

The *reputation update polices and heuristics layer* implements mechanisms to synchronize the two views, and to correlate the newly computed values. For example, a specific function is to trigger an update of $r_{feedback}$, when the, $r_{expected}$, has an unexpected variation, or when its variation trespass some thresholds. A concrete case could be when the volume of sales increases dramatically, with no appropriate variation of $r_{feedback}$. Appropriate heuristics will be presented in the next sections.

It is assumed that the customer and service records follow the model presented in [16] and enhanced in this proposal.

Without losing the generality, but for simplicity of the computation, we adopt the same formula, i.e., (1), as a core mechanism for reputation update engines; only the metrics will be specific to each view. Let us assume that a given service has a starting reputation $r_0$, on both views. We use the formula (2) for computing an updating value of the reputation:

$$r_{current} = r_0 \prod (1 + \lambda_i) \quad (2)$$

where: $\{ \lambda_i = k_i \times w_i \;\; i = 1.....M\}$

M: the number of considered metrics
i: a given i-th metric [I belongs to N]
$k_i$: basic normalized update due to the variation of the i-th metric [$k_i$ belongs to R]
$w_i$: weight factor associated with the i i-th metric [$w_i$ belongs to R]

The *off-line* and *run-time* engines compute the $r_{feedback}$ and $r_{expected}$, respectively, using (2) and appropriate heuristics for adopting $\lambda_i$. In the next section, we introduce a customer dynamic model and show how the $\lambda_i$ are computed.

### 4.1 History metrics

We recall the main concepts of the enhanced recommender model [16], which we consider as the basis for dynamic feedback metrics:

s <r>: each service <s> has an associated reputation <r>
$P_i$<s,$r_i$>: each provider offers a service with its associated reputation
$P_j$<s,$r_j$>: another provider can offer the same service with a different associated reputation
u <e, c>: a user has a credibility and confidence metrics associated with

Note: for simplicity, we consider that <e, c> are the same for any service.

In evaluating the customer feedback, we consider individual metrics and metrics for a class of subscribers. In both cases, the feedback mode, i.e., 'by_request' or 'at_will' helps to differentiate between different extreme feedbacks.

feedbackMode::= {by_request, at_will}

### *4.2 Individual metrics*

For individual metrics, 'subscription seniority', 'feedback timing', and the 'satisfaction' are relevant.

Seniority in profile::= {long term, regular, new}
FeedbackTiming::= {quick, regular, late }
SatisfactionDegree::= { x% | x = 0 - 100}

For policy-specification, we define satisfaction by metrics

satisfaction = satisfied, when $x > \beta_1$
= regular, when $\beta_2 \leq x \leq \beta_1$
= dissatisfied, when $x < \beta_2$

seniority = long term, when term $> \tau_1$
= regular, when $\tau_2 \leq$ term $\leq \tau_1$
= new, when term $< \tau_2$

feedback = quick, when $t < t_2$
= regular, when $t_2 \leq t \leq t_1$
= late, when $t > t_1$

With the above definitions, we assume that the architecture handles the seniority of the subscribers and the timestamps of their feedbacks after a service was consumed.

The following patterns of interest can be identified for each seniority profile:

a. &&<quick><satisfied>]
b. &&<quick><dissatisfied>]
c. &&<late><satisfied>]
d. &&<late><dissatisfied>]

The profile metrics quantified as 'regular' do not alter the computation of the reputation.

With the new metrics, a user is characterized by
- expectation
- credibility
- seniority

and a 'per service' feedback pattern. The feedback patterns comprise:
- feedback timing
- satisfaction degree
- feedback dynamics (#satisfied, #dissatisfied, repetitive replies, observation period)

There is a calibration phase for each system, where the appropriate values are tried and settled for the thresholds. For example, the following steps are considered by a calibration procedure for the feedback timing:

(1) An 'average' reply time is observed and recorded for both feedback modes for a given service.
(2) After the calibration period, the customer reaction is observed for that service, called 'average'.
(3) A policy can be defined by heuristics, as follows:

```
START
IF feedback mode = at_will
   IF 'feedback timing' is 'three times' than the
     'average'
        THEN feedback = late
IFNOT (feedback mode = at_request)
    IF 'feedback timing' is 'twice' than 'average'
        THEN feedback = late
END
```

**Heuristic #1**. Settling the feedback values
In a similar way, and based on calibration, policies for settling each threshold can be defined.

### *4.3 Metrics for classes of subscribers*

For a class of subscribers to the same service, we propose feedback metrics capturing the community behavioral. In the case of a community, the individual profiles are aggregated. In order to capture the dynamicity of the feedback, we introduce a few feedback metrics describing a pattern structure. In a given observation period ($\Delta$), we define the number of repetitive replies (m) and the number of satisfactions (ni+) an dissatisfactions (ni-), as well as n- = max {ni- | i = 1…} and n+ = min {ni+ \ i = 1….}. For example, in Figure 2, on the top, m = 2, n1+ = 4, n2+ = 5, and n+ = 4. In the third basic pattern, when both satisfactions and dissatisfactions are present, a pair (n+, n-) is attached to it.

Based on a series of observations periods, a profiling system is able to classify customers and have a coarse granularity on the feedbacks.

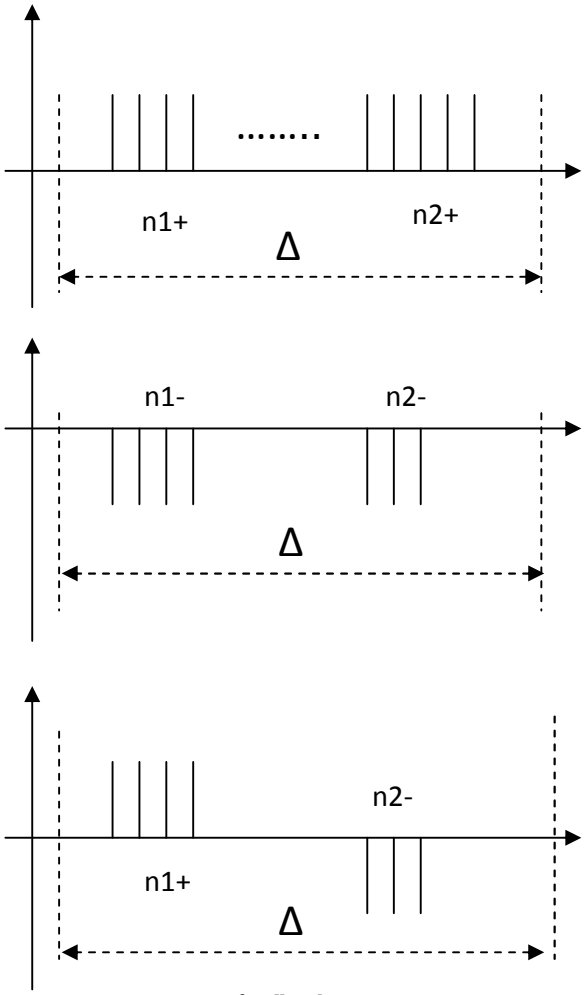



Figure 5. Basic feedback patterns

With these metrics, a reputation engine can trigger appropriate reputation update mechanisms. Definitively, they can be combined with the user history metrics to build more complex (and more accurate) updates.

*4.4 Dynamics on service subscriptions*

Service reputation is a main metric to justify the existence of that service, new investment in developing new features of the service, and new marketing activities to promote a given service.

In our model, we consider a few metrics that portray the dynamics of service sale, e.g., the volume of transactions, the number of new customers in a given period, and the number of lost customers in that period. For the last metric, we also consider the seniority, therefore distinguishing between long term and new customers.

Therefore, apart its features from the quality point of view, from the reputation perspective, a service is described by:

- r: reputation
- vol: variation in transaction volume [± %]
- new: new customers [%]
- lost_new: lost new customers [%]
- lost_long: lost long term customers [%]

(% is considered versus the numbers at the beginning of the observation period; usually every update represents the start of a new period)

A revision of the reputation of a service is always based on the above metrics.

An example of using heuristics for updating the reputation based on the number of transactions can be:

START
IF vol = - 10%
    THEN $r_{real} = (1- 0.1) \times r_{current}$
ELSE
    $r_{real} = r_{current}$
END

Heuristic #2. Updating the reputation versus variations of transactions

*4.5 Conclusion on the reputation update model*

We presented a model for updating the reputation of a service considering the user profile, its dynamic feedback, on the one side, and the dynamics of service subscriptions. In general, the reputation updates is triggered by a significant variation of one of the service subscription dynamics. This moment defines the origin of an updating period. It is assumed that a recommender system records other customer feedback information that is considered when updating the reputation.

In the following section, we present some basic heuristics to update service reputation, considering the metrics described above.

## V. DYNAMIC REPUTATION UPDATING

There are several classes of updates, based on what metrics are used.

*5.1 Satisfaction and feedback based policies*

The simplest way of updating the reputation is considering the first four patterns

a. [<quick><satisfied>]
b. [<quick><dissatisfied>]
c. [<late><satisfied>]
d. [<late><dissatisfied>]

Following Policy #1, we associate with (a) and (b) a correction $\theta_1$ and with (c) and (d) a correction $\theta_2$, with $\theta_2 < \theta_1$, when the feedback_mode is 'at_will' and with $\theta_4 < \theta_3$, respectively, when the feedback_mode is 'by_request', with $\theta_4 < \theta_3 < \theta_2 < \theta_1$. The subjective justification of these values is given by the customer attitude in terms of promptness of reactions and their qualification.

The correction, in this case, is expressed by the following policy [Policy#1]

START
IF feedback_mode = *at_will*
  IF feedback = *quick*
    IF satisfaction = *satisfied*
      THEN $r_{real} = (1+ \theta_1) \times r_{current}$
    IFNOT (satisfaction = *dissatisfied*)
      THEN $r_{real} = (1- \theta_1) \times r_{current}$
  IFNOT (feedback = *late*)
    IF satisfaction = *satisfied*
      THEN $r_{real} = (1+ \theta_2) \times r_{current}$
    IFNOT (satisfaction = *dissatisfied*)
      THEN $r_{real} = (1- \theta_2) \times r_{current}$
IFNOT (feedback_mode = *at_request*)
  IF feedback = *quick*
    IF satisfaction = *satisfied*
      THEN $r_{real} = (1+ \theta_3) \times r_{current}$
    IFNOT (satisfaction = *dissatisfied*)
      THEN $r_{real} = (1- \theta_3) \times r_{current}$
  IFNOT (feedback = *late*)
    IF satisfaction = *satisfied*
      THEN $r_{real} = (1+ \theta_4) \times r_{current}$
    IFNOT (satisfaction = *dissatisfied*)
      THEN $r_{real} = (1- \theta_4) \times r_{current}$
END

Policy #1: Reputation updates #1

The values of all weights are done by validated calibration. For example, if the orders of a given service do not increase (or decrease), it means that the reputation is too high. Therefore, reputation is always 'in question', when the transactions for a service vary, or the service orders show a quick increase or decrease (in terms of volume, and in terms of new customers).

*5.5 Satisfaction, feedback, and seniority based policies*

Let us assume that a mechanism is in place for complying with Policy #1; additionally, the seniority must be considered. There are two strategies used in our model to update the reputation: (i) an optimistic one, and a (ii) pessimistic one.

Let us assume we have the following situation:

vol = +15%
new = 10%
lost_new = 2%
lost_long = 1%

In an optimistic strategy, would credit the 'vol' and 'new', while downplaying the 'lost_new' and 'lost_long'. Following the same approach of correcting by fraction representing the percentage, i.e., 10% is a correction of 0,1, we have the following heuristic:

START
IF vol = +15%
  new = 10%
  lost_new = 2%
  lost_long = 1%
  THEN $r_{real} = (1+ 0{,}15)\ (1 + 0{,}1)\ (1 - 0{,}02)\ (1- 0{,}01) \times r_{Policy\#1}$
END

Heuristic #3. Considering variations in transactions and subscribers (optimistic)

In a pessimistic approach, losing new subscribers or long term subscribers is an indication of service degradation from the quality point of view, of a violation of the SLA with a significant number of subscribers, or simply that the service was not upgraded at the expected standard.

In this case, a multiplicity factor can be used to consider the loss, e.g., k1 for loosing new subscribers and k2 for loosing long term subscribers.

START
IF vol = +15%
  new = 10%
  lost_new = 2%
  lost_long = 1%
  THEN $r_{real} = (1+ 0{,}15)\ (1 + 0{,}1)\ (1 - k_1 \times 0{,}02)\ (1- k_2 \times 0{,}01) \times r_{Policy\#1}$
END

Heuristic #4: Considering variations in transactions and subscribers (pessimistic)

Calibrating the values for k1 and k2 in this case follows also a given heuristic, as expressed below:

| | 1T | 2T | 3T |
|---|---|---|---|
| k1 | 3 | 2 | 1 |
| k2 | 6 | 6 | 6 |

Heuristic #5: Multiplicity correction factors

In Heuristic #5, an example of selecting the multiplicity correction factors is presented. Assume that the unsubscribe event occurs in 1T, 2T or 3T time units for the enrolment, the example gives more weight to the loss of long term customers ($3T < \tau_2$ to correctly evaluate the 'new').

Note1: In the model presented above we considered no difference between the types of service, assuming that the QoS, from the provider perspective was delivered according to the SLA.

Note2: In the heuristics and the metrics presented above, we didn't consider any emotional feedback that might influence the feedback (such as accompanying gifts, bonuses, or penalties for QoS violations), nor particular interests of a customer in a service provider, such as stocks.

*5.2 Reputation update considering feedback patterns*

A fine grain reputation update considers the feedback patterns presented in Figure 5. While only one pattern can be considered to update the reputation, Heuristic #6 considers all three patterns (see Figure 5).

START
IF $(\Delta, m, n^+)$ $[\Delta > \tau_1]$
THEN $r_{real} = (1+ m^+xn^+/100) \times r_{current}$
IF $(\Delta, m, n^-)$ $[\Delta > \tau_1]$
THEN $r_{real} = (1+ m^-xn^-/100) \times r_{current}$
IF $((\Delta, m, n^+, n^-)$ $\&\&\Delta > \tau_1]$
THEN $r_{real} = (1+ m^-xn^-/100) \times (1+ m^+xn^+/100) \times r_{current}$
END

Heuristic #6. Reputation updated considering the feedback patterns

*5.3 Reputation updating policies and heuristics*

By their own nature, from both views (customer, provider), the reputation values on each view is a list, similar with time series, with

$r_{feedback} = \{r1, r2, r3, \ldots\ldots\}$, and (3)

$r_{expected} = \{r'1, r'2, r'3, \ldots\ldots\}$

at {t1, t2, t3, …..}

The reputation values can have the following position, as shown in Figure 6.

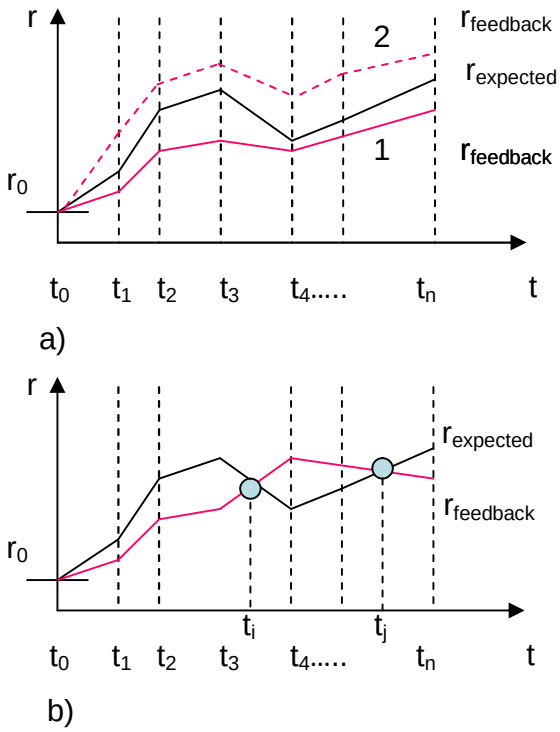



Figure 6. Relative position of reputation values

While computing the real reputation on both views, and considering the dynamics of customer feedback, we can observe a *channel trend* (Figure 6a) or local anomalies (Figure 6b). Formally, there are several cases for defining updating heuristics.

A chann*el trend* is considered when $| r_{expected} - r_{feedback}| < \varepsilon$ and an *anomaly* occurs when $| r_{expected} - r_{feedback}| > \delta$, where $\varepsilon << \delta$, for all $t_i$ ($\varepsilon$ and $\delta$ are thresholds that are established per services).

The model allows implementing five heuristics for synchronizing the feedback and expected reputation (see (3)). This allow to adjust the market from some actions; with no lost generality, we only consider three potential actions, i.e., 'increase/decrease the storage order', 'increase/decrease the price', and 'increase/decrease the offered QoS'. These actions are main contributors for the service costs.

1) Synchronization

In this case, both expected and feedback reputations are in synchronization with small variations.

---

IF
$| r_{expected} - r_{feedback}| < \varepsilon$
AND
$r_i > r'_i$ for all i
OR
$r_i < r'_i$ for all i
THEN "keep same storage order"
AND
"keep same prices"
AND
"keep same QoS/SLA agreements)
END

---

Heuristic #7. Regular computation

No particular actions are triggered, but regular reputation computation by off-line and run-time engines.

2) Pessimistic anomaly (from the provider)

In this case, the feedback reputation is much higher than the expected reputation; it is a situation to increase the objective function (benefits).

---

IF
$| r_{expected} - r_{feedback}| > \delta$
AND
$r_i > r'_i$ for all i
THEN
IF 'no QoS violation"
THEN "increase storage order"
ELSE
"offer lower QoS/SLA provider metrics"
(less costs)
END

---

Heuristic #8. Pessimistic policy

3) Optimistic anomaly (from the provider)

In this case, the expected reputation is much higher than the feedback reputation; it is a situation to decrease the objective function (benefits).

---

IF
$| r_{expected} - r_{feedback}| > \delta$
AND
$r_i < r'_i$ for all i
THEN
IF 'no QoS violation"
THEN "reduce storage order"
OR "reduce price"
ELSE
"offer better QoS/SLA provider metrics"
(more costs, to attract customers)
END

---

Heuristics #9. Optimistic policy

4) Under estimation (by the provider)

This case refers to the situation where the expected reputation decreases, but the feedback reputation increases.

---

IF
$| r_{expected} - r_{feedback}| < \varepsilon$
AND $r_i = r'_i$ for a given I &&see $t_i$ in Fig. 3b]
AND $r_{i-1} < r'_{i-1}$
AND $r_{i+1} > r'_{i+1}$
THEN
(expectation decreases, customer satisfaction increases)
IF 'no QoS violation"
THEN "increase storage order"
OR "increase price"
ELSE
"offer lower QoS/SLA provider metrics"
(less costs))
END

---

Heuristics #10. Under estimation policy

5) Over estimation (by the provider)

This case refers to the situation where the expected reputation increases, but the feedback reputation decreases.

---

IF
$| r_{expected} - r_{feedback}| < \varepsilon$
AND $r_j = r'_j$ for a given I [see $t_j$ in Fig. 6b]
AND $r_{j-1} > r'_{j-1}$

AND $r_{j+1} < r'_{j+1}$
THEN
(expectation increases, customer satisfaction decreases)
IF 'no QoS violation"
THEN "decrease storage order"
OR "decrease price"
ELSE
"offer better QoS/SLA provider metrics"
(more costs)
END

Heuristics #11. Over estimation policy

## VI. A CONTEXT-BASED SIMILARITY MODEL

Then main idea of our approach is (i) having well defined service clusters, (ii) compute the distance between service feature, (iii) evaluate service similarity, based on service features, (iv) consider user-, service,- and producer-based similarity reflected by the appropriate reputations, and (v) evaluate how interchangeable two services are. When a service query is issued the algorithm we propose select the most appropriate service, considering both distance and similarity between services.

### *5.4 Identifying clusters of similar services*

Expanding what was mentioned in [27], service cohesion of a service cluster must be strong (best potential to be similar), while correlation between two service clusters should be weak (service independence). We say that service $s_1$ is similar with $s_2$, and note $s_1 \sim s_2$, if the similarity confidence is greater than a given threshold $\delta$. In a cluster S with $\|S\|$, where $\|x\|$ is the cardinality of x, we redefine cohesion and correlation as follows:

$$Cohe_S = \{(s_i, s_j) \mid s_i \sim s_j (\sim_{thres} > \delta)\} / (\|S\| \times (\|S\| - 1)) \quad (4)$$

and

$$Correl_{S,S'} = (A (S, S') + A (S', S)) / 2 \times \|S\| \times \|S'\|, \quad (5)$$

where

$$A (S, S') = \| \{s_i, s_j \mid s_i \in S, s_i \in S' \; si \sim sj \mid \sim_{thres} > \delta\}\| \quad (6)$$

with $$\sim_{score} = Cohe_S / Correl_{SS'} \quad (7)$$

defining the similarity score.

We notice that $\sim_{score}$ defines similarity classes based on the preexisting service clusters. To enhance the similarity score, clusters aggregation and clusters split operations are possible. Conditions and assessments for doing these are presented in [27].

### *6.2 Distance metrics for service similarity*

Let us assume that a service *s* has *n* features (usually called data-points, as they are expressed by concrete values in an *n* dimensional space). The following distance methods are adapted for comparing services:

(a) Service Euclidian distance between two services in the *n* dimensional space

$$d(s_1,s_2) = 1/n \Sigma (a_{1i} - b_{2i})^{**2}, \text{ for all } i = 1 \ldots n \quad (8)$$

where $a_i$, $b_i$ are service features.

(b) Service city-block distance

$$d(s_1,s_2) = 1/n \Sigma |a_{1i} - b_{2i}|, \text{ for all } i = 1 \ldots n \quad (9)$$

(c) Service Pearson correlation coefficient

$$r(s_1,s_2) = 1/n \Sigma ((a_{1i} - \underline{a})/\sigma_a) \times ((b_{2i} - \underline{b})/\sigma_b), \quad (10)$$

where $\underline{a}$ and $\underline{b}$ are the sample mean of $a_i$ and $b_i$ respectively, and $\sigma_a$ and $\sigma_b$ are the sample standard deviation of $a_i$ and $b_i$.

The service Pearson distance is defined as

$$d(s_1,s_2) = 1 - r(s_1,s_2) \quad (11)$$

(d) Service Cosine similarity

$$d(s_1,s_2) = \cos(\theta) = (s_1 \bullet s_2) / (|s_1| |s_2|) \quad (12)$$

where $\bullet$ is the vector product of $s_1$ ans $s_2$.

By selecting a service distance metric, a clustering algorithm computes the distance matrix between two services. Mostly, (a) and (b) of the above are satisfying the triangle inequality, as true metrics.

### *6.3 Classes of similarities*

In order to select the most appropriate service, we introduce producer-based similarity ($\sim_{prod}$), recommender-based similarity ($\sim_{recc}$), and user-based similarity ($\sim_{user}$). Producer similarity is based on the expectation, recommender's similarity is statistics-based, and user similarity is based on user feedback. In this taxonomy, $s_1 \sim_{prod} \{s_2, s_3, \ldots\}$ define a cluster of similar services, as defined by the producer.

To refine service similarity, we introduce the notions of *primary service features* and *secondary service features*, as shown in Figure 7,

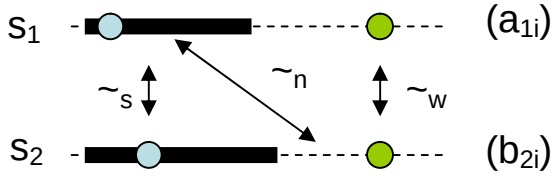



Figure 7. Similarity classes.

where the bold items represent primary service features ($A_1$ set), and the dashed items represent secondary service features ($A_2$ set) (similar for $s_2$)

We introduce strong, weak, and normal similarity, represented by $\sim_s$, $\sim_w$, and $\sim_n$, respectively.

Therefore, ($s_1 \sim s_2$) =

$= \sim_s$, iff all $a_{1i} \in A_1$ and $b_{2i} \in B_1$

$= \sim_w$, iff all $a_{1i} \in A_2$ and $b_{2i} \in B_2$

$= \sim_n$, iff there are $a_{1i} \in A_1$ and $b_{2i} \in B_2$ or
there are $a_{1i} \in A_2$ and $b_{2i} \in B_1$ (13)

Similarity composition allows to capture all possible combinations, e.g,, $\sim_{prod/s}$ represents a strong similarity defined by the producer, based on the primary service features.

A refinement of feature-based similarity is can be expressed when service features do not show a direct semantic matching, but feature composition might lead to such a match. Considering that a subset a service feature for a given service is equivalent with a feature for a service the similarity is computed for, we introduce feature composition-based similarity, as shown in Figure 2.

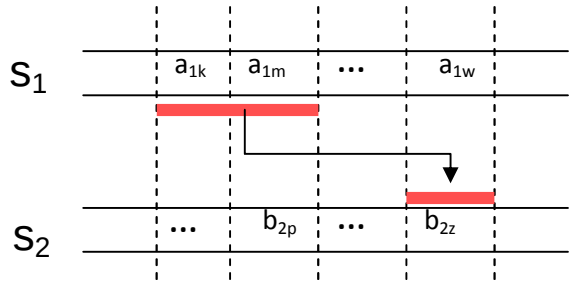



Figure 8. Feature composition-based similarity.

$s_1 \sim_{a1k,a1m\,/\,b2z} s_2$ (14)

with the semantic that the values of $a_{1k}$ and $a_{1m}$ composed are similar to the values of $b_{2z}$. Composition might be any arithmetic or Boolean operator, according to the nature of the features, e.g., if sets, then 'U' (union), if values, then '+' (addition), etc. If type, and $a_{1k}$:T1 and $a_{1m}$:T2, and $b_{2z}$:T3, then, then T3 is a subtype of either T1 or T2.

Combination between $\sim_s$, $\sim_w$, and $\sim_n$, and feature composition-based similarity can be applied following (13).

### 6.4 Updating similarity

When evaluating service similarities, perfect match of service features is desired, but rarely found, due to some continuous values of the features. For example, looking for a service offering the weather temperature with an accuracy of 0.1°F is not feasible. A query on what month the temperature is 67.3F might have no match; but, for a given location, a query on what moth shows [75-80] °F might be answered by April or May, if a Mediterranean area. We identify two possible relaxations when performing the matching.

#### 6.4.1 Context-based feature migration

In time, and based on business models or customer feedback, some primary features become secondary, and vice-versa. Even more, in the same time, in different contexts, a feature can belong to either primary or secondary feature sets.

Let C = {$c_i$} a set of contexts and

$s1 ::= (A_1 \cup A_2)_{context = c1}$, with $A_1 \cap A_2 = \phi$

$s1 ::= (A'_1 \cup A'_2)_{context = c2}$, with $A'_1 \cap A'_2 = \phi$ (15)

then, the following is possible:

$s_1 \sim_{context = c1} s_2$

$s_1 \sim_{context = c2} s_3$ (16)

#### 6.4.2 Feature relaxation-based similarity

Service features are not always perfectly matching (so goes for query matching, as well). Most of the time, the exact matching is not mandatory, e.g., if a service feature has a numeric value a variation of $\alpha_{1i}$ (usually symmetric, but not necessarily) of +/- $\alpha_{1i}$ is allowed. As a result, the similarity metrics presented in II.B can be relaxed. The same relaxation can be applied for similarity on data type/subtype, for similarity concerning the set of interface operations, or similarity concerning variations of an algorithm implementation. For example, when a query (with explicit relaxation of +/- 2ms) targets a service with a response delay of 10ms, any service offering a delay within [8ms, 12ms] is a desired matching. With no explicit relaxation delay, 10ms is mandatory. In this case,

$s1 \sim_{a1i +/- \alpha 1i}$ $s2 <=> b_{2i} \in [a_{1i} - \alpha_{1i}, a_{1i} + \alpha_{1i}]$ (17)

where $a_{1i}$ and $b_{2i}$ are the corresponding features of $s_1$ and $s_2$, respectively.

*6.5 Recommender-based similarity*

Recommender mechanisms rank [22] the products or services based on feedback received after a series of recommendations and successful transactions. The *ranking* is subject to incomplete, fictitious feedback, volume of transactions for a given product or provider, and confidence in feedback. Based on statistics, the recommender computes its own *ranking* per product, defining the reputation (r) of a service/product.

Considering a set of service clusters a recommender build based on type of services/products,, we define:

Cluster = {$cluster_i$}
with $s_1 \in cluster_i$ and $s_2 \in cluster_i$, for a given service feature

$$s_1 \sim_{feature = ai} s_2 ::= |rank_{s1} - rank_{s2}| < \varepsilon_{ai} \quad (18)$$

In general,

$$s_1 \sim_{Uai} s_2 ::= \max \{|rank_{s1} - rank_{s2}|\} < \min \{\varepsilon_{ai}\} \quad (19)$$

*6.6 Customer feedback reputation-based similarity*

Based on customer individual metrics, context, and potential query with relaxation, a reputation is associated with a service/product. Heuristics for updating the reputation have been presented in [22][23]. In general, the following information is available:

s <r>: each service <s> has an associated reputation <r>
$P_i$<s,$r_i$>: each provider offers a service with its associated reputation
$P_j$<s,$r_j$>: another provider can offer the same service with a different associated reputation
u <e, c>: a user *u* has a credibility and confidence metrics associated with
For simplicity, we consider that <e, c> are the same for any service.

For a given user, we define similarity in terms of $r_s$

$$s_1 \sim_{feedback} s_2 ::= |r_{s1} - r_{s2}| < \varepsilon_0 \text{ with } e > e_0 \text{ and } c > c_0 \quad (20)$$

In the following, the newly introduced model is used by an algorithm to identify the most suitable service to satisfy a query for a service.

## VII. ALGORITHM FOR SERVICE RETRIEVAL USING SIMILARITY

We introduced a similarity model and classes of similarity that allow a user (invoker) to use for a service in a given context, allowing or not precise relaxation for some service features, and under different types of similarity (strong, weak, normal). Distance metrics were also adopted for services, in order to cluster the most suitable services for a particular query, before computing the similarity.

Based on the model previously introduced and on the user model [23] and reputation [22], a query for a service *s* can be expressed as

**Q** (s, similarity type, context, with/without relaxation on {$a_{1i}$})

The algorithm presented below illustrates the main steps to reach a service proposal that can be a set, a given service, or no service at all.

*Algorithm for finding a requested service* query **Q**, *based on similarity between potential satisfying services*

---

1: **begin**
2: **identify** the service cluster &&&see (4)]
3: **select** a distance metric &&&see (5)-(9)]
4: **calculate** distance between all $s_i$ in the cluster
5: **select** a subset {$s_k$ with min {$d(s_i, s_j) < \varepsilon$}
6: **if** Q with relaxation
7: **apply** (10) and (11) for all mentioned features
8: **if not**
9: **if** Q with context
10: **apply** (12) and (13)
11: **if not**
12: **compute** a subset {si} of the set found before step12
13: **select** {sl} from the subset of step 12, with rank ($s_l$) > $\delta_1$ and $r_{feedbakc} > \delta_2$ &&&see (16) and (17)]
14: **select** a subset for the subset of step 13
15: **return** the subset of step 14
16: **end**

---

Note that the output of step 15 might be an empty set, or a set having many recommended services complying to the query conditions.

The complexity of the algorithm is given mostly by the number of services features that can be considered with relaxations.

A variation of the algorithm was experimented with relaxation conditions for a set of contexts. The number of features with relaxation, the number of contexts, and the number of services into a cluster determine the performance of the algorithm.

Different experiments on the on-line Barnes&Nobles system (on-line book shopping) show a reasonable improvement on the precision the algorithm returns after running various numbers of queries and varying different conditions.

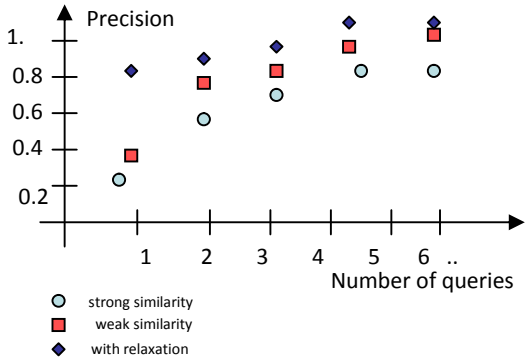



Figure 9. Precision of service returned to queries with different types of similarities

With no surprise, a service satisfying a query with relaxation riches quicker and with a higher precision the query expectation.

*6.8 Similarity issues*

We presented an approach for service invocation using similarity taxonomy where weak, strong, and normal similarity. Practically, services are clustered and service distance/similarity metrics were adopted from text-based domains. A reputation-based mechanism (introduced in [22][23]) is used in combination to context-based similarity and feature relaxation methods to identify a set of services that better serves a given query.

We also introduced the techniques of feature aggregation when similarity is evaluated and the continuous update of feature classification, i.e., primary/secondary, according to the context. More work should be done on these two items, as semantic-based aggregation should be considered.

*6.7 Feature relaxation-based similarity*

Service features are not always perfectly matching (so goes for query matching, as well). Most of the time, the exact matching is not mandatory, or, at least, the query can explicitly mention an acceptable variation. Usually, this is expressed as a constraint associated with a given service feature. For example, requiring a book delivery service, might have as a condition, *delivery costs* < *threshold*. In other cases, if a service feature has a numeric value a variation of $\alpha_{1i}$ (usually symmetric, but not necessarily) of +/- $\alpha_{1i}$ is allowed. As a result, the similarity metrics presented in II.B can be relaxed. The same relaxation can be applied for similarity on data type/subtype, for similarity concerning the set of interface operations, or similarity concerning variations of an algorithm implementation.

For $s_1$ [$a_1$, $a_2$, $a_3$,….$a_n$]
$s_2$ [$b_1$, $b_2$, $b_3$,…$b_k$]

Let us assume that a few service features $a_i$ are associated with constraints. These constraints may be expressed as follows:

$a_i$ > expression/threshold
$a_i$ < expression/threshold
$a_i$ ϵ &&&x, y] (*belongs to*, as an interval)
ai ϵ {x, y} (*belongs to*, as a set)

For a service selection, all expressions must be returned TRUE.

A query for a service can be represented by:

Q (s, similarity type, context, {($a_i$, constraint$_i$)})

We express this as

$s_1 \sim_{constraint} s_2$ <=> bi satisfies $a_i$, and all constraint$_i$ are evaluated TRUE, for ALL *i* mentioned in the Q

For example, when a query (with explicit relaxation of +/- 2ms) targets a service with a response delay of 10ms, any service offering a delay within [8ms, 12ms] is a desired matching. With no explicit relaxation delay, 10ms is mandatory. In this case,

$$s1 \sim_{a1i +/- \alpha 1i} s2 <=> b_{2i} \epsilon [a_{1i} - \alpha_{1i}, a_{1i} + \alpha_{1i}] \quad (21)$$

where $a_{1i}$ and $b_{2i}$ are the corresponding features of $s_1$ and $s_2$, respectively.

As a note, similarity with constraints increases the chance to have a matching to a given query, on the expense of additional computation. A variation of this kind of similarity is when constraints are:

- Mandatory, for primary features (M)
- Optional (while desired), for secondary features (O)

For expressing these variations, a Q must be explicit on the categories of features

Q (s, similarity type, context, M:{($a_i$, $constraint_i$)}, O:{($a_i$, $constraint_i$)})

A response for the system should also contain the reference to the kind of feature/similarity, e.g.,

A response can be

$\{s_{1/M/nonO}, s_{3/M/O}, s_{8/M/nonO}\}$, or simply

$\{s_{1/M/nonO}, s_3, s8/M_{/nonO}\}$,

where *nonO* index represents the feedback of not all optional constraints were evaluated TRUE.

With no surprise, a service satisfying a query with relaxation riches quicker and with a higher precision the query expectation.

Two performance improving procedures are possible: (i) Building a query cluster each query belongs to (as a set of a priori known services that might satisfy the mandatory features of a given query, and (ii) Based on the responses, it is interesting that the same similarity criteria used to identify similar services can be used to evaluate the 'satisfaction' of the returned services.

The first procedure needs a look up in the previous query inventory, and group them by context and user. Then, looking by context and user ID, a subset of services are derived. A Q is now answered by considering a particular service cluster, sensitively smaller than the entire set of services.

By running a few situations, with 1,000 services, 5,000 contexts, 8,000 users, and queries with 4 mandatory features and 5 optional features, no constraints, the following results were obtained (Figure 10).

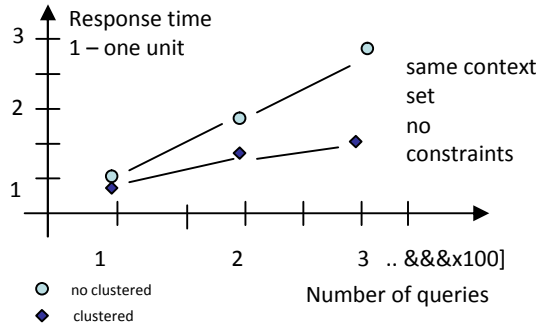



Figure 10. Response time versus number of queries in clustered and non clustered approaches

In the case where services are clustered (based on previous answers) the response time is dropping by almost half. However, a similarity (Q, A) must be also executed to evaluate hoe the recommended services satisfy a given query.

In terms of returned services, under the same setting, the results are presented in Figure 11.

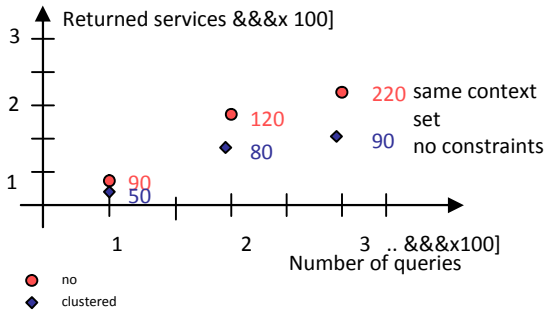



Figure 11. Returned services versus number of queries in clustered and non clustered approaches

A sensitive drop in returned services is observed in the clustered mode; however, it seems to be saturation with the number of queries increasing. This can be caused by the lack on context differentiation, or by the similarities of the queries. On the latter part, more experiments are needed.

The second procedure is used to apply similarity (Q, A) for each service in the retuned set, and select that service that has the max similarity. The procedure is simple, but requires to be applied, in turn, to all returned services. To simplify, one may select to run the similarity check only for the primary features. In this case, with the same settings for the experiment, the computation time is reduced by two thirds. This is due to the fact that similarity with constraints requires additional computation for each feature to validate that the constraint is TRUE.

To substantially reduce the computation time, and the cardinality of the returned services, a condition on service reputation reduced the time and the number of returned services.

## VIII. CONCLUSION AND FUTURE WORKS

The paper presented a framework and appropriate mechanisms to evaluate the services/providers in the light of their respective direct impact on user perception. Essentially, the proposal considers several innovative ways of considering user impact on an accurate evaluation of a service/provider reputation.

The proposed schema can capture indirect service delivery and allow reputation correction based on the real transactions.

Future investigations should focus on a more formal definition of service/provider/feature similarity and the stability of the reputation accuracy over a longer period. This might lead to the reputation predictions; specialized metrics for assessing the accuracy of predictions in the light of indirect delivery are challenging but seen as very helpful in web-service driven environment.

On the user side, consistency feedback and reliability should be correlated with the frequency of users' report and transaction peaks, as well as with the user's report patterns. This will allow detection of potential 'off-market' agreements between providers and set an appropriate service level agreement policy.

One aspect that is left out, but worth to be mentioned, is that the reputation adaption function presented in the thesis should be refined. We adopted a linear $r = r_0 (1 + \lambda)$ reputation adaptation function. However, some services might listen to other forms of reputation adaptation function, as $r = r_0 (1 + e^{\lambda/r0})$, or $r = r_0 (1 + \log\lambda/r_0)$, etc. We think that this can be approached by defining on the customer side and on the provider side, a most suitable function for reputation update, considering the type of service and the context.

Another aspect that should be validated with more date sets is related to the duration of the trying period for endorsing the expected reputation. For example, 1-2 moth for a book service, 1-2 weeks for a coffee service, or 6 months for a piece of software. In this case, trying various validation periods might provide a more accurate reputation update.

Service reputation was calculated 'per context'. A global view, form the producer perspective will require studies on weighted cross-context reputation, in the case of a free-context query. For example, a formula might be

$$S_{cross\text{-}context} = (\Sigma\, w_i \times r_{contexti}) / \Sigma w_i \qquad (22)$$

In this case, large statistics are needed for an optimal tuning of $w_i$.

Another aspect that was surprising concerned the fact that new customers where quiet close to the estimated reputation; this raise the issue of a finer tuning considering customer profiles. The heuristics presented might need updates considering exceptions.

Finally, the fact that some service features shown strong impact on service reputation were originally classified as secondary (or, even miscellaneous) requires a more customer-oriented service design. Actually, the conclusion is quite the opposite with what is going on with the service launching, when a myriad of features are attached to a service, without a clear evaluation of a need. Even more, new features are added, without accurate validation of the use and customer evaluation of the existing ones.

## IX. REFERENCES

[1] P. Resnick and H. R. Varian. Recommender systems. Communications of the ACM, 40(3):56–58, 1997.

[2] Esma Aimeur and Flavien Serge Mani Onana. Better control on recommender systems. E-Commerce Technology, 2006. The 8th IEEE International Conference on and Enterprise Computing, E-Commerce, and E-Services, The 3rd IEEE International Conference on Volume, Issue , 26-29 June 2006 pp. 38 – 38

[3] T. D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. Communications of the ACM, 35(12):61–70, 1992.

[4] Dieberger, A., Dourish, P., Hook, K., Resnick, P., and Wexelblat, A. Social navigation: techniques for building more usable system. Interactions, 7, 6, 2000, 36-45.

[5] Adomavicius, G. and Tuzhilin, A., Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Transactions on Knowledge and Data Engineering, Volume 17, No. 6, June 2005. Pp. 734 – 749

[6] Chen, Y. B., and Xie, J. H. Online consumer reviews: a new element of marketing communications mix. Working paper, Eller College of Management, University of Arizona, Tucson, AZ, 2004.

[7] Kumar, N., and Benbasat, I. The influence of recommendations and consumer reviews on evaluations of Web sites. Information Systems Research, 17, 4, 2006, 425-439.

[8] Dini,O., Moh, M., Clemm, A.: Web Services: Self-adaptable Trust Mechanisms. AICT/SAPIR/ELETE 2005: 83-89

[9] Massa, P., & Bhattacharjee, B. (2004). Using Trust in Recommender Systems: An Experimental Analysis. Trust Management, (Proceedings of the 2nd International Conference, iTrust 2004).Oxford, UK, LNCS 2995, Springer, pp. 221-235.

[10] M. Pazzani, "A Framework for Collaborative, Content-Based, and Demographic Filtering, Artificial Intelligence Rev., pp. 393-408, Dec. 1999.

[11] A.I. Schein, A. Popescu, L.H. Ungar, and D.M. Pennock, Methods and Metrics for Cold-Start Recommendations," Proc. 25th Ann. Int'l ACM SIGIR Conf., 2002.

[12] R. Jurca and B. Faltings. "An Incentive Compatible Reputation Mechanism", Proc. IEEE Conf. on E-Commerce, pp. 285-292, Newport Beach, CA, June 2003.

[13] Pei-Yu Chen, Yen-Chun Chou, Kauffman, R.J. Community-Based Recommender Systems: Analyzing Business Models from a Systems Operator's Perspective. Proceedings of the 42nd Hawaii International Conference on System Sciences (HICSS-42). pp. 1-10, January 2009.

[14] Kwei-Jay Lin, Haiyin Lu, Tao Yu, and Chia-en Tai. A reputation and trust management broker framework for Web applications. Proceedings of The 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service. April 2005. pp. 262- 269

[15] Dellarocas, C. Strategic manipulation of Internet opinion forums: implications for consumers and firms. Mgmt. Sci., 52, 10, 2006, 1577-1593.

[16] O. Dini, P. Lorenz, and H. Guyennet; An Enhanced Architecture for Web Recommenders, SERVICE COMPUTATION 2009, IEEE Press, pp.

[17] Mean opinion score and metrics, http://technet.microsoft.com/en-us/library/bb894481.aspx [accessed: Jan 10, 2010]

[18] Adomavicius, G. Tuzhilin, A. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Transactions on Knowledge and Data Engineering, Volume 17, No. 6, June 2005. Pp. 734 – 749

[19] PESQ, PESQ Algorithm firmeware, http://www.qoesystems.com/products/pesq_basic.htm &&acessed: January 13, 2010]

[20] Bruckner, R. M., List, B. and Schiefer, J., Striving Towards Near Real-Time Data Integration for Data Warehouses, In Proc. of the 4th Intl. Conf. on Data Warehousing and Knowledge Discovery (DaWaK 2002), Springer LNCS 2454, pp. 317–326, Aix-en-Provence, France, Sept. 2002.

[21] Schiefer, J., Seufert, A. 2005. Management and Controlling of Time-Sensitive Business Processes with Sense & Respond. In Proceedings of International Conference on Computational Intelligence for Modelling Control and Automation. Vienna, Austria.

[22] O. Dini, P. Lorenz, and H. Guyennet; An Enhanced Architecture for Web Recommenders, SERVICE COMPUTATION 2009, IEEE Press, pp. 372 – 378, ISBN: 978-1-4244-5166-1, Athens, Greece

[23] O. Dini, P. Lorenz, A. Abouaissa, and H. Guyennet, Dynamic Feedback for Service Reputation Updates, ICAS 2010, pp. 168-175 ISBN: 978-1-4244-5915-5, Cancun, Mexico

[24] C. Wu and E. Chang, Searching Services 'in the web': A Public Web Services Discovery Approach, SITIS 2007, The Third IEEE Conference on SignalImage Technologies and Internet-based Systems, pp. 321-328.

[25] M. Paolucci, B. Shishedjiev, Xh. Zenuni, and B. Raufi, GHSOM-based Web Service Discovery, 2010 European Computing Conference, ISSN: 1790-5117, 2010

[26] M. Szomszor, C. Cattuto, H. Alani, K. O'Hara, A. Baldassarri, V. Loreto, and V. D. Servedio, "Folksonomies, the semantic web, and movie recommendation," In 4th European Semantic Web Conference, Bridging the Gap between Semantic Web and Web 2.0, 2007.

[27] X. Dong, A. Halevy, J. Madhavan, E. Nemes, and J. Zhang, Similarity Search for Web Services, The 30th VLDB Conference, Toronto, 2004

[28] S. Cost and S. Salzberg, A Weighted Nearest Neighbor Algorithm for Learning Symbolic Features. Machine Learning, No. 10, 1993, pp. 57-78

[29] L.S. Larkey and W. Croft, Combining Classifiers in text Classifications Techniques, ACM SIGIR 1998.

[30] H.-H. Do and E. Rahm, COMA – A System for flexible Combination of Schema Matching Approaches, VLDB 2002

[31] A.M. Zaremski and J.M. Wing, Specification matching of software components. TOSEM, No. 6, pp. 333-369, 1997

[32] C. Bouras and V. Tsogkas, Improving text summarization using noun retrieval techniques, LNCS, Knowledge-based Intelligent Information and Engineering Systems, vol. 5178/2008, pp. 593-600

# Sharing Building Information with Smart-M3

Kary Främling
Aalto University
PO Box 15500, Espoo, Finland
e-mail: Kary.Framling@hut.fi

André Kaustell
Åbo Akademi University
Joukahaisenkatu 3-5, FIN-20500 Turku, Finland
e-mail: andre.kaustell@abo.fi

Ian Oliver
Nokia Mobile Solutions - Platforms
Helsinki, Finland.
e-mail: Ian.Oliver@nokia.com

Jan Nyman
Electrical Building Services Centre
Posintra Oy
Kipinätie 1, FIN-06150 Porvoo, Finland.
e-mail: jan.nyman@posintra.fi

Jukka Honkola
Nokia Research
Helsinki, Finland.
e-mail: Jukka.Honkola@nokia.com

*Abstract*— **Semantic nets are a universal information structure that can be used for representing nearly any kind of information. This is why the semantic web has also chosen to use them as the universal format for representing data, usually using RDF (Resource Description Framework) as the syntax. Semantic nets are also suitable for sharing information between different domains, organizations, manufacturers etc. In this paper, we describe how a semantic net and agent-based shared storage called Smart-M3 has been implemented and can be used for such information sharing. The particular application domain studied is building automation, where interoperability between equipment made by different manufacturers is rare. This is a great challenge for implementing "ubiquitously smart buildings", where building automation systems, user interfaces and services could interact. The paper describes how the Smart-M3 concept can be used as an enabler of interoperability, where an ecosystem of supplementary services is created through manufacturer-agnostic agents.**

***Keywords - Smart Buildings, Smart-M3, semantic net, ontologies, software agents.***

## I. INTRODUCTION

Creating smart buildings and smart environments in general has been a topic of research and development for a long time. However, such environments are still largely found only in experimental or pilot environments despite their potential to make people's lives easier, reduce energy consumption and environmental footprint, as well as improve the quality of life in general. In this paper, we describe a distributed information architecture that makes it possible to implement such smart environments on a large scale by integrating information access to and control of different building automation systems. We also show how smart buildings can be created as parts of smart environments.

Building automation is a domain where interoperability is a challenge due to conflicting interface and communication standards, e.g. KNX, LON, Modbus etc., in addition to a great number of prorietary solutions. Solutions to these interoperability challenges have been developed e.g. in the ongoing DIEM project (Devices and Interoperability Ecosystem, http://www.diem.fi) using device and protocol adapters that enable unified information access to them all on the Internet Protocol level and notably through Service Oriented Architecture (SOA) solutions. Such SOA-based solutions are good in the case where standards (real or de-facto) exist for the semantic representation of the information. In practice, there is a lack of universal standards. Meanwhile there tends to be many potential interfaces available developed by different organisations and projects, which are not interoperable. This lack of compatibility is a major obstacle for creating *Smart Spaces* where humans and devices could interact smoothly [1].

The Smart Spaces notation is heavily overloaded and has been used for describing a wide variety of things. In this paper, we use it to signify a geographical space where information is available about the space itself, the devices and services available in it, the people present in it and about other potentially useful information or services. Such a Smart Space concept has been initially proposed in [2] as a solution to enabling interoperability. As no standards exist that would cover the information representation needs of such Smart Spaces, we believe an incremental process will occur [3]. In the first phase, devices and systems will publish their available information and services using their current semantic notations (standardized or not). When the information becomes available, that makes it possible to

create new services that use the information, while augmenting it with information about the services themselves and information produced by them.

In the Sedvice/M3 architecture, information is expressed as subject-relation-object Triples that build up labeled, directed multi-graphs (one or more). In the rest of this paper we will call such graphs *semantic nets* even though graph theory and semantic net theory use partially different vocabularies and present other potential incompatibilities due to their background and history. The Triples are represented using Resource Description Framework (RDF) notations. Triples are stored and managed by the Semantic Information Broker (SIB), which can be distributed over many devices. The Smart Space Access Protocol (SSAP) is used for performing operations on the semantic net.

After this introduction, the paper gives a state of the art overview of building automation systems, semantic nets and Smart Spaces. In Section III we describe the whole system architecture and in Section IV we show the current level of implementation, followed by conclusions.

## II. BACKGROUND

Building automation systems and Smart Spaces are currently two distinct domains with different technological and scientific backgrounds, which is the reason for splitting this section. We will provide an overview of the state-of-the-art for both domains, as well as some background for the work reported in this paper.

### A. Building Automation Systems

Systems integration in buildings has traditionally been about physical dimensions, voltage, plug dimensions etc. Control mechanisms usually control either one device only (e.g. a lamp, a refrigerator etc.) or power supply for security reasons (e.g. fuses, main switch etc.). Implementing integrated functions such as switching the power off from certain appliances, cutting off water supply and activating the burglar alarm with one single "leaving home" command has required a lot of dedicated cabling and custom devices, installed by professionals.

Different communication standards have been defined in order to provide more feasible solutions, such as LON (http://www.lonmark.org/), KNX (http://www.knx.org/) and ModBus (http://www.modbus.org/). However, none of these has become a global standard that all manufacturers would support. Many solutions based on these protocols also tend to be expensive to install, maintain and upgrade. Furthermore, they are not conceived in a way that would allow for easy integration between them; in fact, they may even on purpose be designed in a way that makes interoperability more difficult due to commercial reasons.

Meanwhile, remote monitoring and control of buildings has become a common functionality at least for bigger buildings such as shopping centers, office buildings, libraries etc. Remote monitoring services are becoming an increasingly important part of the business of traditional building companies as well as other companies. These systems tend to use internet as the information channel because it is cheap to set up and use. As people become increasingly connected to the internet from their homes, internet and the communication protocols associated with it have become an interesting option also for building automation solutions. The fact that many multimedia devices (including mobile phones) integrate internet connectivity by default, makes it possible to take systems integration and usability to levels that are not possible with "classical" building automation systems.

Figure 1 illustrates how different devices can be connected to a "protocol converter" that makes device information available through internet protocols. The "protocol converter" can be an ordinary computer or a cheaper and more energy-efficient solution, such as the Home Control Center (http://smarthomepartnering.com/cms/) initially proposed by Nokia. Device connectivity is implemented through adapters that convert the underlying protocols into a generic internet interface.

In practice, a generic internet interface nowadays signifies a browser-compatible format (HTML and others) for user interfaces and XML messages for machine-readable information. For successful machine-to-machine communication, the semantics of the XML messages have to be understood in the same way by both parties. The currently most used method for describing message semantics is XML Schemas. In building automation, the oBIX (Open Building Information Xchange, http://www.obix.org/) is an example of such a protocol. Devices Profile for Web Services (DPWS) is another initiative with similar goals. In addition to these, more generic messaging protocols exist that are intended for communication with any kind of devices (not only related to building automation). The PROMISE Messaging Interface (PMI) [4] is an example of such an interface. The IP for Smart Objects (IPSO) alliance (www.ipso-alliance.org) has similar objectives but it is unclear whether they have yet specified any messaging protocols. In practice, none of these has obtained global acceptance.

There are still technical and functional differences between the protocols. Currently, oBIX and PMI are the easiest protocols to compare because their specifications are readily available. For the moment, oBIX does not support real-time events due to the lack of callback functionality, which is included in PMI. On the other hand, oBIX is more well-known. oBIX is also REST-compliant [5], whereas PMI currently lacks the functionality of accessing resources (e.g. devices, sensor values etc.) directly via a URL. The lacking features would be easy to add both for oBIX and PMI but for the moment especially the lacking features in oBIX make it hard or impossible to implement some functionality that is essential in real-life applications. As a conclusion, no universal standard currently exists for representing information neither about smart objects in general, nor about building automation systems.


Figure 1. Connecting devices to the Internet.

### B. *Semantic nets for describing Thing-related information*

Buildings are per definition complex products for which an extensive amount of documentation is produced both during the design and the manufacturing phases. CAD models and other technical information produced during the design phase are some of the most essential parts of the product information available when the building is taken into use. CAD models are a form of semantic networks that explicitly model "part-of", "depends-on" and similar relationships. The bill-of-materials (BOM) used during manufacturing is another important piece of product information that can be represented as a semantic network. However, the BOM is usually representing product information on a product-type or product-variant level rather than on a product-item level. The classical BOM is not sufficient for managing building product information, where each building is an individual product-item, where even the parts of the building have individual properties and where parts can be changed during the product lifecycle [6].

Semantic networks represent sets of named relationships between different nodes (or objects) in a network. Using a collection of pair-wise relations between nodes, where every relation may also have an associated "strength", can represent a semantic network. Relation strengths are particularly useful when semantic networks are used for reasoning, e.g. for diagnostic or prognostic purposes as those needed in many middle-of-life (MOL), i.e. in-usage applications of product items [7].

Solutions for managing semantic networks in a multi-organizational context are being developed under the name "semantic web". RDF and Web Ontology Language (OWL) are examples of standards being developed for the semantic web. Software frameworks also exist that use these standards, e.g. Jena (http://jena.sourceforge.net/), OpenRDF (http://openrdf.org/) and the Redland RDF Application Framework (http://librdf.org/).

However, RDF and OWL are mainly focused on describing web content rather than on describing product information. Furthermore, the related software tools are not, as such, designed to be used for implementing distributed applications. Therefore agent frameworks could be more suitable for this purpose. Examples of such agent frameworks are ABLE (http://www.alphaworks.ibm.com/tech/able) and JADE (\http://jade.tilab.com/) that integrate inter-organizational communication. In a multi-agent framework, agent references correspond to links between nodes of a semantic network. Therefore agent frameworks could be used as building blocks for a distributed implementation of semantic networks for describing product information.

When using an agent framework with support for data analysis and decision support, the nodes themselves can also be "intelligent". Especially the data analysis methods included in the ABLE framework could be applicable as decision support systems. ABLE data analysis and decision support agents provide support for many different data analysis methods, e.g. naïve Bayes, decision trees and neural networks. In addition to these, ABLE agents exist for both crisp and fuzzy rules that are useful for explicitly expressing expert knowledge. This portfolio of methods is particularly interesting for MOL scenarios that include diagnostics, prognostics and condition-based maintenance. ABLE agents can be trained both on- and offline and included in different software components to perform filtering or decision-making on different levels.

The Design Pattern [8] concept developed in the context of Object-Oriented Programming is an example of how object relationships, which are conceptually quite identical to semantic relations, can be combined with processing in well-documented ways that are known to be "good" from a program design point of view. In the DIALOG software platform [9], semantic relations have been used as a way of storing product-related information structures, while agents implemented the needed information processing in order to apply some major Design Patterns to the domain of product lifecycle information management [10].

DIALOG has been used for real-life applications in many application domains, such as shipment tracking (inventory management, detecting delays, project management), building automation (intelligent refrigerator, remote monitoring, condition-based maintenance), automotive (online tracking and collection of sensor and other data, condition-based maintenance) and telecommunications (configuration management). There are also ongoing projects in the same and new domains. Semantic nets have proven their value for storing Thing-related information, while the agent concept is a flexible and efficient paradigm for the processing of that information. However, DIALOG is implemented using "traditional" database and networking technologies, which are not initially conceived for semantic net and agent-based information processing. Smart-M3 is a paradigm and platform developed to overcome those limitations.

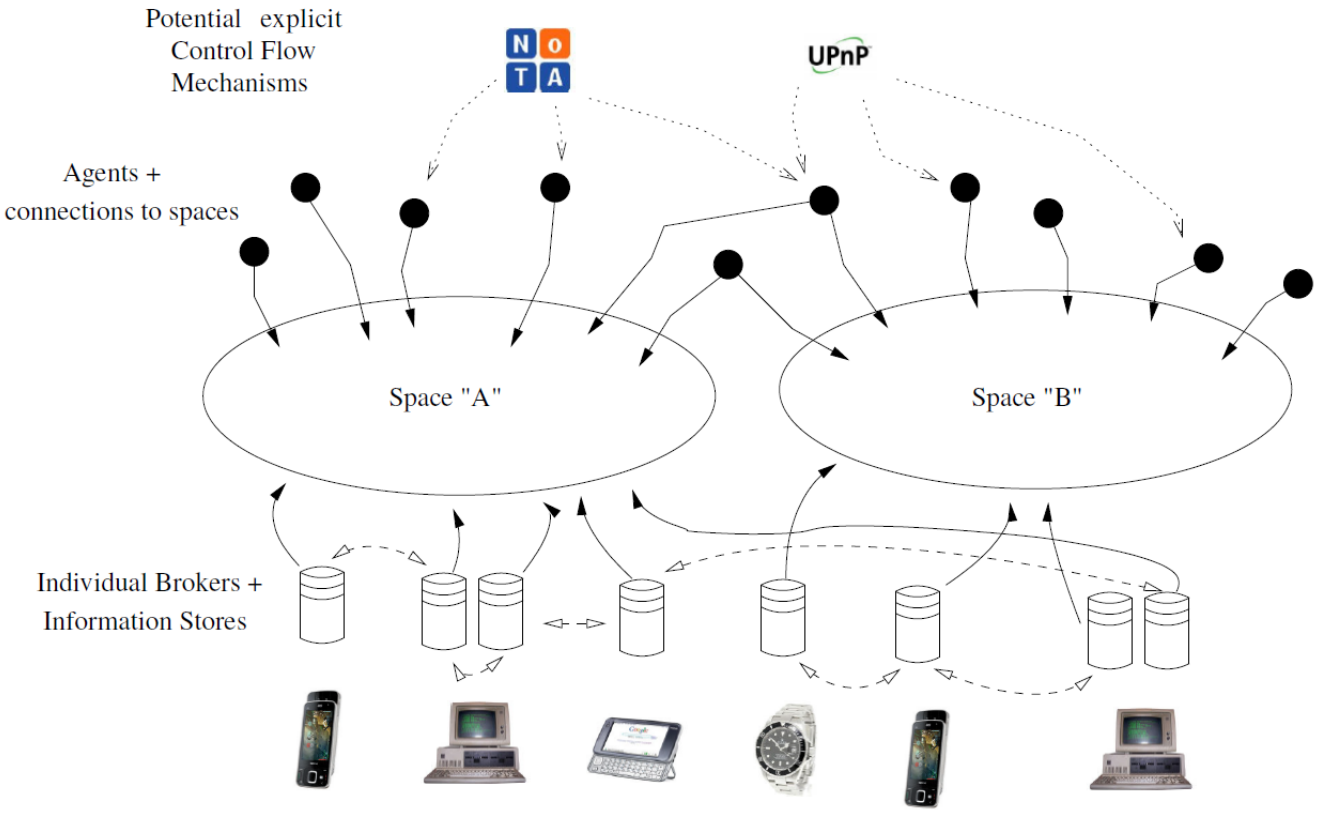



Figure 2. High-level system architecture and connectivity to external systems.

### C. *Semantic Net and Agent-based information processing with Smart-M3*

The Smart-M3 system [2][11] consists of a space based communication mechanism [12][13] for independent agents. The agents communicate implicitly by inserting information to the space and querying the information in the space. The space is represented by one or more Semantic Information Brokers (SIBs), which store the information as an RDF graph. The agents can access the space by connecting to any of the SIBs making up the space by whatever connectivity mechanisms the SIBs offer. Usually, the connection will be over some network, and the agents will be running on various devices. The information in the space is the union of the information contained in the participating SIBs. Thus, the agent sees the same information content regardless of the SIB to which it is connected. The high-level system architecture is shown in Figure 2, which includes the distribution routing between SIBs and external interfaces to protocols such as NoTA and UPnP from the agents.

The agents may use five different operations to access the information stored in the space:

*Insert*: Insert information in the space
*Remove*: Remove information from the space
*Update*: Atomically update the information, i.e. a combination of insert and remove executed atomically
*Query*: Query for information in the space
*Subscribe*: Set up a persistent query in the space; changes to the query results are reported to the subscriber

In addition to these access operations there are *Join* and *Leave* operations. An agent must have *joined* the space in order to access the information in the space. The join and leave operations can thus be used to provide access control and encrypted sessions, though the exact mechanisms for these are still undefined.

In its basic form the M3 space does not restrict the structure or semantics of the information in any way. Thus, we do not enforce nor guarantee adherence to any specific ontologies, neither do we provide any complex reasoning[1] [14][15]. Furthermore, information consistency is not guaranteed. The agents accessing the space are free to interpret the information in whatever way they want.

We are planning to provide, though, a mechanism to attach agents directly to the SIBs. These agents have a more powerful interface to access the information and can be e.g. guaranteed exclusive access to the information for series of operations. Such agents may perform more complex reasoning, for example ontology repair or translation between different ontologies. However, they may not join any other spaces but are fixed to a single SIB and thus a single space.

[1] The current implementation of the concept understands the owl:sameAs concept

The M3 spaces are of local and dynamic nature, in contrast to semantic web which embodies Tim Berners-Lee's idea of semantic web [16] as a "giant global graph". We envision that the spaces will store very dynamic context information, which poses different challenges than the internet-wide semantic web. For example, in order to provide a true interoperability for local ubiquitous agents, the space (i.e. SIBs) will have to provide a multitude of connectivity options in addition to http: plain TCP/IP, NoTA [17], Bluetooth, RFID [18]. Furthermore, the space should be fairly responsive. While we do not aim for real-time or near real-time systems, we think response times need to remain within seconds in order to be acceptable.

The responsiveness is one of the factors behind the fundamental decision to not enforce any specific ontologies and allowing the agents to interpret the information freely, as it lessens the computational burden of the infrastructure. Another, and more important reason is that we explicitly want to allow mashing up information from different domains in whatever way the agents see best. Strict ontology enforcement would make this kind of activity extremely difficult as all new ways of mashing up the information would require approval from some ontology governance committee. However, we still plan to provide means for ontology enforcement for cases where the space provider explicitly wishes to restrict the ways the information is. Such situations will occur in reality where such enforcement is the best approach.

The information content in a M3 space may be distributed over several SIBs. The distribution mechanism assumes that the set of SIBs forming a M3 space are totally routable but not necessarily totally connected. The information content that the agents see is the same regardless of the SIB where they are connected [19]. Distribution may also occur between first order space interaction as described in [20].

Security is provided firstly as an effect of the localised nature of spaces coupled with the agent-join mechanisms. Within the space there is need for a more sophisticated policy mechanism to regulate access, update and the trust of the information at both individual triple and larger RDF graph structure levels [21].

### *D. Applications in M3 Spaces*

The notion of application in M3 space differs radically from the traditional notion of a monolithic application. Rather, as a long term vision, we see the applications as possible scenarios which are enabled by certain sets of agents [22][23][24]. Thus, we do not see an email application running in M3 space, but we could have a collection of distributed agents present which allow for sending, receiving, composing and reading email. Figure 3 pictorially depicts the relationship between the user, her agents and, in this case, one space, while Figure 4 shows the user (via agents) interacting with many spaces.

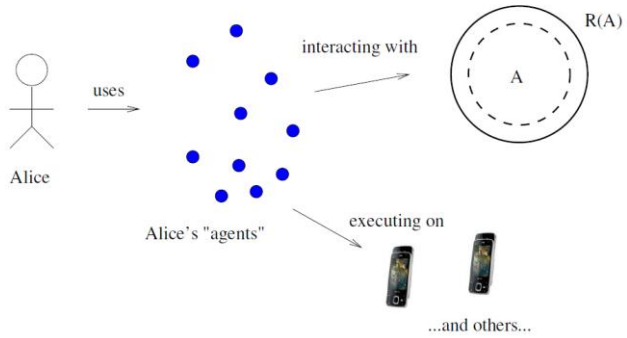



Figure 3. A User's Agents, Devices, Spaces and Information.

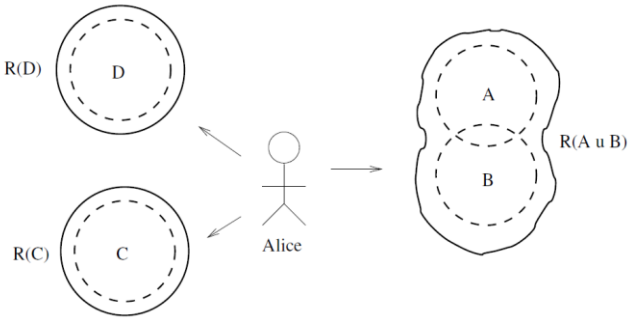



Figure 4. A User and Multiple Spaces

For this kind of scenario based notion of application, we also would like to know whether the available agents can successfully execute the scenario. The envisioned model of using this system is that the user has a set of agents which are capable of executing certain scenarios. If a user needs to perform a new scenario that the current set of agents are not capable of executing, she could go and find a suitable agent from some directory by describing the desired scenario and the agents she already has.

Thus, we need some formal or semi-formal way of describing agent behavior both with respect to the M3 space and to the environment. While there exists research addressing behavior in multi-agent systems, for example by Herlea, Jonker, Treur and Wijngaards [25], this kind of ad-hoc assembly of agents in order to execute a certain scenario seems to be quite unaddressed in current research. However, slightly similar problems have been addressed in e.g. web service orchestration research [26], but these still seem to concentrate on design-time analysis rather than run-time analysis. As for shorter term, our vision is that sets of existing applications would be enhanced by being able to interoperate and thus allow execution of (automatic) scenarios that would have been impossible or required extensive work to implement without the M3 approach.

## III. BULDING SPACE AND SERVICES IT CAN PROVIDE

Despite the lack of universally accepted ontologies for representing information related to buildings and the systems found in them, the application domain still presents some advantages [27]. It is possible to identify a common name for some key concepts, such as "temperature" and "humidity".

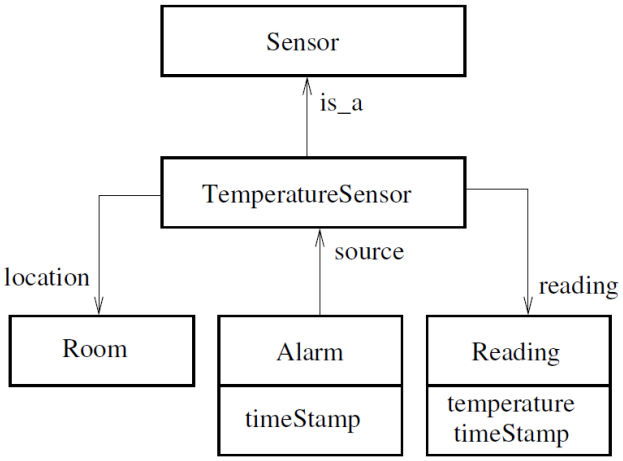



Figure 5. Example of an Ontology (Schema).

When it comes to the $CO_2$ level it already becomes more difficult to agree on a common name. There may also be several different sensors of the same type. For instance, a ventilation machine with heat recovery would normally have at least four temperature sensors that need to be named. Still, the number of manufacturers of such machines is typically not too big (less than ten in Finland) so even though all manufacturers would choose their own names for those sensors, it could still remain manageable. We also believe that as equipment manufacturers start publishing information in the smart space and there are services built upon that information, there may also be more incentive for the manufacturers to start using common ontologies such as the simple example in Figure 5 written in a UML or Entity-Relationship style notation. Another approach could be to use "forced semantics" [28] implemented by "translation agents" that would automatically translate information from one ontology to another [29].

Figure 6 shows an example of a small semantic net [30] expressed as an RDF graph for representing sensor values from three different temperature sensors, of which two are located in the same room. This graph satisfies the ontology given in Figure 5, where the reader should be able to figure out the names of relationships (other than object typing) which are not shown for clarity.

This limited net also illustrates some basic processing needs, implemented by agents. In Figure 6, sensor *ts3* has produced three temperature readings, which is the beginning of a reading history. With an increasing number of sensors that may store historical information in the space, it becomes necessary to at least implement *cleaning agents* who take care of removing expired information or removing the "least useful" information if memory is filling up. To avoid losing too much information when cleaning, *summarization agents* become essential. Summarization agents will keep track of minimum, maximum, running average etc. values even after the cleaning agents have removed the original values.

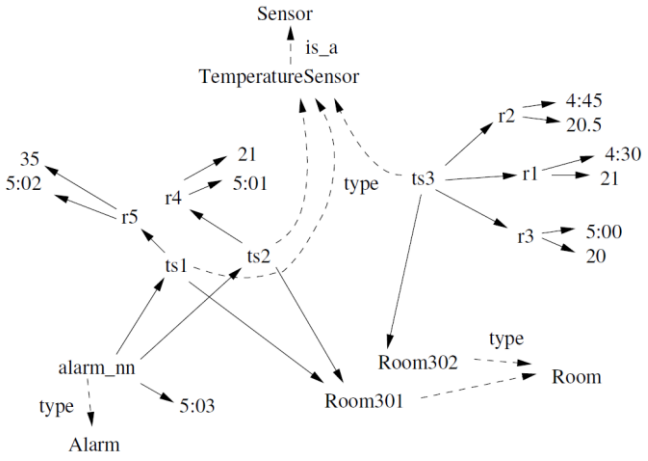



Figure 6. Example of partial semantic net for a home with several temperature sensors.

Using Figure 6 we can write queries across this graph to obtain readings such as those described above. We nominally use here a graph traversal language such as WQL [31] or XPath - M3 specifically supports WQL at this time and a SPARQL parser is being developed.

Given a specific temperature sensor (*ts2*), the query to obtain the current temperature would take the pseudo-code form:

```
ts2 | readings.filter(
      latest(timestamp).
      temperature
```

Given a specfic room, the average temperature would take the form:

```
Room | ( location-1.readings.
       temperature.asBag() )
       / size(location-1.readings)
```

where the suffix -1 denotes inverse traversal of a link and the functions `latest()`, `asBag()`, and `size()` take their common sense meanings when working with sets (or bags) of values.

Figure 6 also shows the alarm event *alarm_nn* in the space as an example of how to handle anomaly detection. A sensor consistency check agent notices an abnormally great value difference for sensors *ts1* and *ts2* that are in the same room. The presence of a new alarm event can be detected by a user notification agent that takes care of notifying the user about the situation. The user can then take the appropriate action, after which the alarm event is removed manually or automatically when the anomaly is no longer present.

It is quite easy to imagine a great number of other functionality that agents could implement based on the information in the space. Such functionality could be deciding on the target temperature in a room based on some "voting" rule, automatically telling the coffee and tea making machines how many people prefer which one, automatic agreement of the next appointment based on the calendar information available from participants in a meeting etc. However, the implementation of such functionality is more

complex than the anomaly detection described previously. It is also much less certain to what degree users want to have or even accept such functionality. However, the purpose of this paper is not to claim that some specific service or functionality is useful as such, the objective is rather to show that Smart-M3 significantly simplify the creation of such services.

Finally, Smart Spaces as described here are not constrained to buildings. They can also cover greater geographical areas and be dedicated for other application domains. One example of such a domain would be publishing weather information collected from private weather stations, ventilation machines etc. for improved local weather forecasts, thereby improved control of heating and cooling in buildings and, as a consequence, improved energy-efficiency as a whole.

## IV. IMPLEMENTATION

The Electrical Building Services Centre of Posintra Oy in Porvoo, Finland, develops demonstration solutions as well as commercially applicable components for integrated building automation. The developed systems should be low-cost, easy to use and easy to integrate with existing and new building automation systems. Current systems installed both in the Centre's demonstration facility and in real buildings make it possible to bridge between the Internet, and a number of building automation protocols and proprietary device protocols. The software platform is based upon OpenWRT (http://www.openwrt.org), a Linux software distribution for embedded systems. The platform makes it possible to communicate with proprietary building automation protocols, and translate the messages to a common format, namely oBIX. This makes it possible to network the devices together, which previously were isolated from each other because of the lack of an universal protocol. The platform itself is running on inexpensive consumer grade hardware (a wireless router with Universal Serial Bus). Currently supported systems include real-time energy metering, data collection and control of air handling units, a control unit for electrical systems of small buildings, wireless power outlets, a consumer-end weather station etc.

The difference of this approach compared to earlier attempts to create a common building automation protocol, is that we have a protocol-agnostic approach. Any building automation protocol can be integrated to the platform by means of adapter software and hardware, and thus we can enable any protocol for Internet connectivity. By connecting together various building automation protocols, we make it possible to combine the functionalities of various subsystems, and create new services that would not be possible without seamless integration of the subsystems. Currently, the subsystems are combined together by the oBIX protocol, which makes it possible to build hierarchal systems by interconnecting the devices on a local level and export the combined information to upper-level systems via oBIX as illustrated in Figure 7.

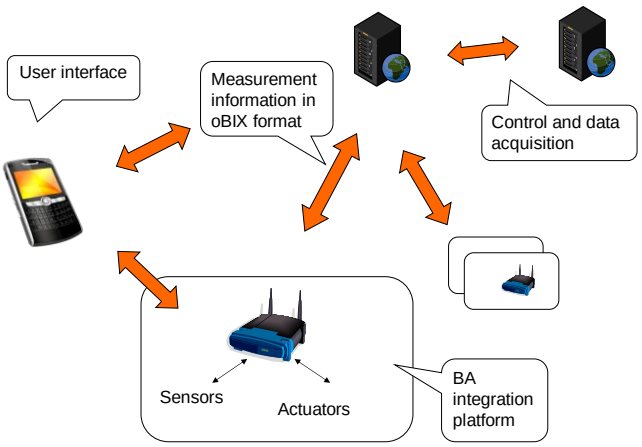



Figure 7. Integration platform makes it possible to control and acquire data from building automation systems, and export the data to back-end systems.

Converting data from proprietary protocols to oBIX simplifies the creation of traditional Supervisory Control and Data Acquistion (SCADA) applications, because the SCADA application now needs to understand only one protocol, instead of a myriad of protocols.

However, optimal control of a building's automation system also needs information from other sources than the various systems located in the building. A simple example is to use weather forecast information from a meteorological website so that the control system can decide to start heating the house during the night (when electricity is cheaper) if the weather forecast says the next day will be colder. Including this type of information from outside the domain of building automation is difficult, if we have to use a building automation specific data format like oBIX.

The Smart-M3 architecture is being integrated to the demonstration platform to make it possible to combine information from various data sources, and to do automated reasoning over it as illustrated in Figure 8. Reasoning agents might change over time, or be only temporarily available, e.g. if they are located in a visitor's mobile phone, PDA or similar.

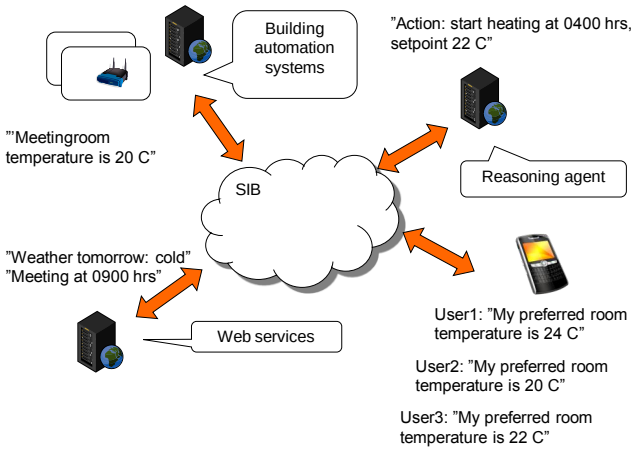



Figure 8. The SIB is an implementation of a data store supporting reasoning over cross-domain information.

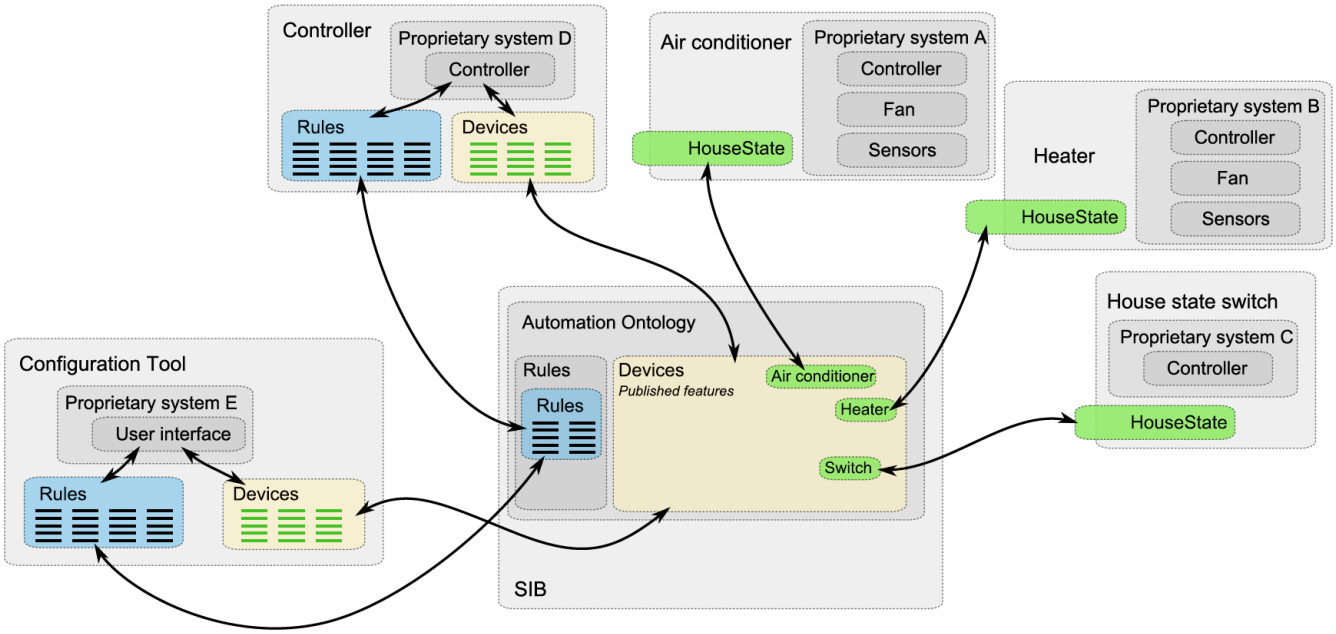



Figure 9. Case study overview showing an interoperability solution for a simple building automation problem.

The integration to the SIB is not intended to replace the control logic embedded in building automation systems, but rather to facilitate using the information from the building automation systems together with other information.

### A. *Smart-M3 Implemented for Building Automation*

In building automation there are several different automated devices which benefit from the knowledge of three simple states. These states are "Home", "Away" and "Vacation". This state data can be modified by status changes in several different system, and this data could be requested by any device compatible with the Smart-M3 stack, running an agent built for this purpose. This simple case study was expanded to be configurable to demonstrate an additional benefit which could be added by the Smart Space approach. Our demonstration scenario requires a home state switch, reflecting the global state (i.e. "Home", "Away" and "Vacation"), and a heating system. In addition to these devices there are two additional parts for enabling interoperability: a controller and a configuration tool. In addition to the required interoperating components there is also a temperature display, and a temperature slider which can be configured to correspond to or set the different temperatures available from the heater appliance. The demonstration application consists of several agents [2] representing the functionality of the devices and a user interface. A conceptual model is shown in Figure 9.

All these components contain a proprietary solution, and provide only a limited set of services through their agents. Additional services could be added to the model, and published to the SIB in order for the configuration tool to make new rules of interaction. The demonstration implementation contains a temperature service concept in addition to the house state concept shown in Figure 9. The temperature data service is contained in the Heater, Temperature Slider and the display. An example configuration is to set the display to show the active temperature setting in the heater, but if there was an independent temperature sensor it could be configured to display its value as well.

The configuration tool can add and remove dependencies, rules and connections by querying the SIB. In our scenario the heater and air-conditioning agents can be configured to request changes in the State switch value, or remove their dependency. The state is indicated by an integer value. When configuration has been set up, the configuration tool agent can leave the Smart Space, as all the necessary data is contained within the ontology in the SIB.

#### 1) *Example Scenario*

All devices in the home connect to the SIB, through their respective SIB interfaces, and insert information about themselves. This information consists of a user friendly name, a list of services it provides and the data which describes its state. No automatic configuration about how they interact exists at this time. When the configuration tool is run, the user is presented with devices registered in the SIB, and can then configure rules.

Rules are interpreted by a controller agent. The controller subscribes to changes in the data of the devices. In order to catch changes in state of the switch the controller listens to new instances of the `Event` class. This instance contains information about what has occurred. When the controller receives a new instance of an `Event` it parses through the list of rules, and if there is a matching rule, it will execute the rule. In this simple implementation a matching rule will create a new instance of the class `Invoke` and add properties to it according to the configured rules.

[2] Smart-M3 agents are also called *Knowledge Processors*.

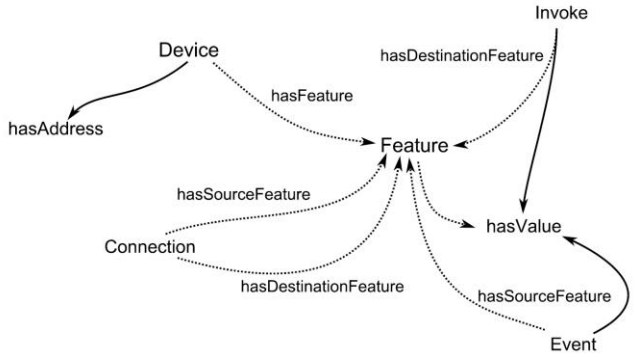



Figure 10. Overview of the ontology used in the implementation of the case study.

The new `Invoke` instance is subscribed to by the agent representing the service invoked, and can then be used to alter the internal state accordingly.

*2) Ontology*

In this use-case we created an ontology containing rules for automation, and concepts for expressing house state, and temperature. The house state is mapped to an integer value, but should more correctly be defined as an enumerated class consisting of the three states as individuals 'Home", "Away", "Vacation". An overview of the ontology is shown in Figure 10. In this way the configuration tool could suggest potentially interesting counterparts for creating connections or rules. The ontology does not contain a complete set of concepts for building automation, but merely the concepts needed for this specific automation task and demonstration. The ontology is expressed in OWL-DL and contains classes, data properties and object properties.

*3) Code Generation*

The Python Code Generator was used to generate the agent ontology API. We recognized several features which would be required in order to create a practical implementation of the initial plan of the building automation ontology. Updating data in the SIB requires a delete and insert query. There is no support for subscribing to changes in properties, and thus the implementation uses classes of `Event` and `Invoke` to register changes. This approach adds significant overhead to network utilization. The generated API also populates all instance properties depending on which class it is loaded from, which could also be changed to a populate on demand approach in order to reduce unnecessary queries to the SIB, or alternatively it could make use of a better query receiving all properties in one message. At the moment all properties are queried separately.

*4) Building Automation Configuration Tool*

The tool for creating rules for the controller is a command line tool. The configuration tool inserts or removes instances of `Connection`. Typing "?" or "help" provides the list of commands and a brief explanation. There are three main commands; `list`, `connect` and `disconnect`.

With the list command a list of registered devices are shown. The list also shows rules for the controller, as shown in the following listing of configuration tool output:

```
HomeStateSwitch (addr=17)
   0 State
      This     three     state     switch
      corresponds   to   the   Home,   Away,
      Vacation     modes     used     in
      heaters etc.
      Connected to:
         Heater => State
```

*5) Running the demonstration*

The source code is available from SourceForge (www.sourceforge.net), under the name "smart-m3". The demonstration is tested only for specific versions of the components, but may well work with other versions as well. If you are having trouble running the demonstration, please consider installing the following versions; Python 2.6.x, PyQt v.4.5.4 for Python 2.6 and Nokia SIB revision 98. Python is needed because this generated API and the Smart-M3 Mediator[3] are written in Python. The Qt library is used for the message pump of the persistent agent. The Nokia SIB provides the database back-end and connection library for the Smart-M3 Mediator. The SIB is also available from SourceForge, with installation instructions. To the best of our knowledge it does not compile on Windows.

The agents in the building automation demonstration try to connect to a smart space named "x" on 127.0.0.1 at port 10010 by default, but this can be changed for all agents in the file `SmartSpaceConf.py`. Port 10010 is the default for the SIB, but the smart space name must be provided. Running `python SIB.py x` starts a SIB running locally at port 10010 with the smart space name `x`.

The simplest use-case to run is the HomeStateSwitch and the Heater. They have pre-configured addresses of 17 and 7 respectively. These can be connected by the configuration tool using the following commands:

```
Command] list
  HomeStateSwitch (addr=17)
    0 State
        Connected to:
          -
Command] connect
Source address: 17
Source feature: 0
Destination address: 7
Destination  feature:  0  (State  might  have
another number)
Rule name: TestRule
```

If these commands have executed correctly the heater will output its changing state following the home state switch.

The agents can be started in any order, but the configuration tool agent does not find any devices until they are started. The suggested order is to run the controller and the configuration tool, and then any of the service providing devices/agents. Observe that subscriptions to the SIB result

[3] A caching middleware for accessing the SIB.

in a TCP timeout if no subscribed data is sent by the SIB within a quite short period of time depending on network configuration. The Python platform cannot independently set TCP keep-alive messages. It is therefore recommended to run the demonstration locally. There are some subscriptions which are not required after running the configuration tool, thus the example might work even after this timeout error.

The SIB does not clear all data when running the `clear` command, and thus it is recommended to remove the smart space file if problems occur. This is done by running `quit` in the SIB command line interface, and then `rm x`, where `x` is the smart space name, and then running the SIB again `python SIB.py x`.

*6) Lessons learned*

*Tool-chain*: We used the Application Development Kit, ADK, available from SourceForge. The ADK was able to generate the Python API from the use-case's OWL ontology without problems. It was found that there are several features to be implemented into the ADK which would improve this building automation use-case. The ADK does not support subscriptions to changes in properties. This lead to modifications in the ontology in order to use subscription to instances instead. An improved ontology API would reduce overhead of creating new instances of `Event` for changes in values of the devices and sensors, instead of updating one property.

*Smart-M3*: The architecture of the interoperability package as described is not, in its current form, very well suited for building automation. However, it raises interesting points about potential cross domain interoperability scenarios, which are much easier to implement as a result of access to data in a well structured form. By revealing an interface to the smart space for programmatically accessing features in devices, normally accessed by infrared remote controls and proprietary systems, the possibility of interoperating between the virtual and the physical world is realized. Cross domain implementations can be aided by the ADK, by loading several ontologies for a single agent and using input data from one domain and translate it into another.

There is significant overhead in communication and application size. To the best of our knowledge the current SIB implementation does not scale beyond a very modest number of devices in a building automation scenario, and cannot automatically be distributed over multiple SIBs. The workaround to a distributed environment would be to implement an agent which moves data between two SIBs. We recognize that the overhead is partially due to the ADK generated ontology API and this specific implementation.

*Improvement proposals*: The implementation could be further improved by removing the addressing scheme used here. It should not be needed as long as the instances are uniquely identifiable. The solution used here is for convenience when querying for a specific device, we need only an integer value instead of the UUID. The current implementation is still inadequate for real building automation applications, but demonstrates a working concept of connecting features together via the SIB.

## V. CONCLUSIONS

Buildings are a major context for creating smart environments and achieving the goals of Ubiquitous Computing, where people could interact seamlessly with their everyday environment and where the various devices in the environment co-operate in order to achieve some "smart" behavior. However, there are still great interoperability challenges between systems in the domain of building automation due to several competing bus standards and proprietary solutions. The paper shows how such issues can be solved using device adapters and protocol conversion in so called home gateways as illustrated in Fig. 1.

However, the question of semantic interoperability is not solved by home gateways. Semantic interoperability can be partially achieved by standards based on e.g. XML Schema such as oBIX and PMI but it does not seem probable that such standards will achieve a similar global acceptance as HTTP and HTML in any near future. Furthermore, those standards do not define device-specific semantics, such as the names of devices, sensors, alarms etc. In order to overcome such limitations, this paper describes an information publishing mechanism that does not require any pre-defined standard for making the information visible, discoverable and usable to others. That also signifies that it becomes possible for third-party solution providers, who are not themselves manufacturers of building material or building automation, to create Smart Space applications. Such solution providers can provide agents or agent frameworks that implement new functionality. Therefore, our goal is to provide an easy to use basic mechanism that makes it possible to create an open ecosystem where the set of potential applications is open and impossible to predict in advance.

The home gateway solutions described in the paper are currently in use in many real buildings and are expected to become commercial-level volume products within a year. The Smart-M3 implementation described in the paper is implemented on a demonstration laboratory level and will eventually be tested in real pilot targets. Earlier experiences using semantic nets and agents with "classical" tools such as the DIALOG platform have shown their power in several domains such as shipment tracking, product lifecycle management etc. The technical, conceptual and business feasibility of Smart-M3 as an enabler of semantic nets and agents in the building automation domain still remains to be proven. However, the ad hoc data and processing distribution mechanisms of Smart-M3 that are conceived also for embedded devices, and notably mobile phones, are expected to be key enablers of future smart environments where buildings, vehicles, public spaces etc. can be accessed and used in a uniform way.

## ACKNOWLEDGMENT

This work has been carried out in the Devices and Interoperability Ecosystem (DIEM) project (www.diem.fi), financed by the Technology Development center of Finland (TEKES) and by the partner organizations of DIEM, and in

the AsEMo project financed by the European Regional Development Fund.

## REFERENCES

[1] K. Främling, I. Oliver, J. Nyman, and J. Honkola, "Smart spaces for ubiquitously smart buildings," Proc. Third International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM), Sliema, Malta, October 11-16 2009, pp. 295–300.

[2] I. Oliver and J. Honkola, "Sedvice: A triple space computing exploration environment," Proc. Tripcom workshop, April, 2008.

[3] D. Lewis, Convention: a philosophical study. Blackwell Publishing, 2002, 0-631-23257-5.

[4] PROMISE, "Volume 3: Architecture reference: Promise messaging interface (pmi),". [Online; accessed 25 August 2010] http://cl2m.com/system/files/private/PROMISE AS Volume 3 Architecture Reference PMI.pdf, 2008,.

[5] R. T. Fielding, "Architectural styles and the design of networkbased software architectures," Ph.D. dissertation, University of California, Irvine, 2000. [Online; accessed 21 January 2011] http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm.

[6] K. Främling, T. Ala-Risku, M. Kärkkäinen, and J. Holmström, "Agent-based model for managing composite product information," Computers in Industry, vol. 57, no. 1, 2006, pp. 72-81.

[7] K. Främling and L. Rabe, "Enriching product information during the product lifecycle," Proc. 12th IFAC Symposium on Information Control Problems in Manufacturing (INCOM), 17-19 May 2006, Saint-Etienne, France, 2006, pp. 861–866.

[8] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, Design Patterns – Elements of Reusable Object-Oriented Software. Addison Wesley, Reading, MA, 1995.

[9] DIALOG, "Distributed information architectures for collaborative logistics," 2001. [Online; accessed 21 January 2011] http://dialog.hut.fi/.

[10] K. Främling, T. Ala-Risku, M. Kärkkäinen, and J. Holmström, "Design patterns for managing product life cycle information," Communications of the ACM, vol. 50, no. 6, 2007, pp. 75–79.

[11] I. Oliver, "Information spaces as a basis for personalising the semantic web," Proc. 11th International Conference on Enterprise Information Systems, May 2009.

[12] L. J. B. Nixon, E. Simperl, R. Krummenacher, and F. Martin-Recuerda, "Tuplespace-based computing for the semantic web: a survey of the state-of-the-art," The Knowledge Engineering Review, vol. 23, no. 2, June 2008, pp. 181–212.

[13] B. Hayes-Roth, "A blackboard architecture for control," Artif. Intell., vol. 26, no. 3, 1985, pp. 251–321.

[14] A. Passant, ":me owl:sameas flickr:33669349@n00," Proc. Linked Data on the Web (LDOW 2008), Beijing, China, April 2008.

[15] K. Idehen and O. Erling, "Linked data spaces and data portability," Proc. Linked Data on the Web (LDOW 2008), Beijing, China, April 2008.

[16] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," Scientific American, May 2001.

[17] "Network on terminal architecture," http://www.notaworld.org, 11 2008.

[18] J. Jantunen, I. Oliver, S. Boldyrev, and J. Honkola, "Agent/spacebased computing and rf memory tag interaction," Proc. 3rd International Workshop on RFID Technology - Concepts, Applications, Challenges (IWRT 2009), May 2009.

[19] S. Boldyrev, I. Oliver, and J. Honkola, "A mechanism for managing and distributing information and queries in a smart space environment," Proc. 1st International Workshop on Managing Data with Mobile Devices (MDMD 2009) 6-7 May, 2009 - Milan, Italy, 2009.

[20] I. Oliver and S. Boldyrev, "Operations on spaces of information," in Proc. IEEE Conference on Semantic Computation. Berkeley, CA., September 2009.

[21] A. Toninelli, R. Montanari, L. Kagal, and O. Lassila, "Proteus: A semantic context-aware adaptive policy model," in POLICY. IEEE Computer Society, 2007, pp. 129–140.

[22] I. Oliver, E. Nuutila, and S. Törmä, "Context gathering in meetings: Business processes meet the agents and the semantic web," Proc. 4th International Workshop on Technologies for Context-Aware Business Process Management (TCoB 2009), May 2009.

[23] J. Honkola, H. Laine, R. Brown, and I. Oliver, "Cross-domain interoperability: a case study," Lecture Notes in Computer Science, vol. 5764, Springer Berlin / Heidelberg, September 2009, pp. 22-31.

[24] S. Balandin, I. Oliver, and S. Boldyrev, "Distributed architecture of a professional social network on top of m3 smart space solution made in pcs and mobile devices friendly manner," Proc. Ubicomm 2009. Malta, 2009.

[25] D. E. Herlea, C. M. Jonker, J. Treur, and N. J. E. Wijngaards, Multi-Agent System Engineering, ser. LNCS. Springer, 1999, vol. 1647, ch. Specification of Behavioural Requirements within Compositional Multi-agent System Design, pp. 8–27.

[26] H. Foster, S. Uchitel, J. Magee, and J. Kramer, "Compatibility verification for web service choreography," Proc. IEEE International Conference on Web Services, July 6–9 2004. IEEE, 2004, pp. 738–741.

[27] H.-G. Kim and Anseo-Dong, "Pragmatics of the semantic web," Proc. Semantic Web Workshop. Hawaii, USA, 2002.

[28] S. Staab, "Emergent semantics," IEEE Intelligent Systems, vol. 17, no. 1, 2002, pp. 78–86.

[29] B. Hu, S. Dasmahapatra, P. H. Lewis, and N. Shadbolt, "On capturing semantics in ontology mapping," Proc. AAAI. AAAI Press, 2007, pp. 311–316.

[30] M. A. Rodriguez and J. Bollen, "Modeling computations in a semantic network," CoRR, vol. abs/0706.0022, 2007.

[31] O. Lassila, "Programming Semantic Web Applications: A Synthesis of Knowledge Representation and Semi-Structured Data," Ph.D. dissertation, Helsinki University of Technology, November 2007.

# Unscented Transform-based Dual Adaptive Control for Mobile Robots: Comparative Analysis and Experimental Validation

Marvin K. Bugeja and Simon G. Fabri
Department of Systems and Control Engineering
University of Malta
Msida, Malta
Email: marvin.bugeja@um.edu.mt, simon.fabri@um.edu.mt

***Abstract*—Adaptive control involves both estimation and control, which are generally interdependent and partly in conflict. Yet, the majority of adaptive controllers separate the two by assuming that certainty equivalence holds, even if this is not the case. In contrast a *dual adaptive* controller, based on the idea postulated by A. Fel'dbaum in the early 1960s, aims to strike a balance between estimation and control at all times. In this manner, the control law is a function of the estimates' uncertainty, besides the estimates themselves, thereby leading to improved control performance. Few such controllers have ever been implemented and tested in practice, especially within the context of intelligent control, and to the best of our knowledge none on mobile robots. This paper present two novel dual adaptive neural control schemes for the dynamic control of mobile robots in the presence of functional uncertainty. Furthermore, by means of realistic Monte Carlo simulations and real-life experiments, a thorough comparative analysis is performed. A notable novel contribution of this work is the use of the unscented transform within the context of dual adaptive control, aimed at improving further the performance of the system.**

***Index Terms*—Dual adaptive control; nonlinear stochastic control; neural networks; unscented transform; mobile robots.**

## I. Introduction

A major motive for adaptive control is the need to have automatic systems that operate satisfactorily in the ambience of uncertainty. The uncertainty is typically due to unknown and/or time-varying structure or parameters pertaining to the system or process under control. Hence, in addition to keeping the controlled variable tracking its reference, an adaptive controller needs to simultaneously estimate the unknown system functions or parameters. These two objectives, termed *control* and *estimation* respectively, are generally interdependent and partly in conflict, in that typically estimation improves with perturbing (persistently exciting) input signals, while tracking performance does not. On the other hand, good tracking performance still requires good estimates.

Most of the adaptive controllers proposed over the past fifty-five years, including the well-established model-reference adaptive systems (MRAS) and self-tuning regulators (STRs), artificially separate estimation and control via the heuristic certainty equivalence (HCE) assumption. In this manner the parameter estimates are used in the control law as if they were the true values of the unknown parameters, without any due consideration to their inherent uncertainty. Though simple to implement, and adequately applied in many applications, HCE adaptive control can lead to large tracking errors and excessive control actions, which can excite unmodelled dynamics or even lead to instability and possibly physical damage [1]. These effects are more pronounced in situations characterized by high uncertainty, short control horizon and/or time-varying system parameters [2], [3].

This work was supported by National RTDI under Grant RTDI-2004-026.

The issue of simultaneous estimation and control is best addressed via stochastic adaptive control theory. Unlike deterministic approaches, in stochastic adaptive control the uncertainty in the system; be it due to unknown process parameters, noisy measurements, or both; is characterized by probability distributions and their associated statistical measures. Consequently, the whole system is described via a stochastic dynamic model, and the simultaneous estimation and control problem boils down to the minimization of the expected value of a pre-specified cost function. However, this task is rarely straightforward and the general conditions guaranteeing the existence of an optimal control scheme are yet unknown [2].

A major contribution in the field of stochastic adaptive control was made by A. A. Fel'dbaum in his seminal work on optimal control [4]–[6]. Fel'dbaum postulated that the control signal of an optimal adaptive system should have dual goals, namely: (i) to ensure that the controlled variable tracks the desired reference signal, with due consideration given to the estimates' uncertainty, and (ii) to perturb the plant sufficiently so as to accelerate estimation, thereby reducing quickly the uncertainty in future estimates. These two properties are commonly known as *caution* and *probing* respectively, or in Fel'dbaum's own terminology *directing* and *investigating*. Controllers exhibiting these features are named *dual adaptive*. In contrast to an HCE controller, a dual adaptive control law is dependent on the estimates' uncertainty, besides the estimates themselves, and aims to strike a balance between estimation and control at all times. Fel'dbaum also showed that the exact solution to the optimal adaptive dual control problem can be derived using *dynamic programming*, specifically by solving

the Bellman equation. However, in almost all practical situations, with the exception of a few very simple examples [7], the Bellman equation is impossible to solve, both analytically or numerically, due to the very large dimensions of the underlying space [2], [3], [8], [9].

The difficulty in finding the optimal dual adaptive control law in almost every practical case, led to the development of a number of simplified approaches, that though suboptimal, still exhibit the dual properties of caution and probing featured by the optimal dual solution. These suboptimal dual adaptive control schemes can be coherently divided into two groups, namely *implicit* and *explicit* methods. Implicit solutions try to introduce approximations to render the Bellman equation tractable [10], while explicit solutions reformulate the problem via modified cost functions that explicitly include a term related to parameter estimation, in order to induce a form of probing [8], [9], [11]. As pointed out on several occasions [8], [9], implicit solutions are typically more complex and more computationally intensive.

Dual adaptive control has been applied successfully in a number of practical applications [12]–[15]. However none of these applications involve mobile robots. Motion control of mobile robots has captured the interest of numerous researchers over the past three decades [1], [16]–[26]. This interest stems from a vast array of existing and potential practical applications [27]–[31], as well as from a number of particularly interesting theoretical challenges enriching this field of study. In particular, due to their mechanical configuration most wheeled mobile robots (WMRs) manifest restricted mobility, giving rise to nonholonomic constraints in their kinematics. Moreover, many of these WMRs are also underactuated since they exhibit less control inputs than degrees of freedom. Consequently, the linearized kinematic model of these robots lacks controllability, full-state feedback linearization is out of reach [18], and pure smooth time-invariant feedback stabilization of the Cartesian model is unattainable [32].

Most of the early contributions in the field of WMR motion control focus solely on the kinematic/steering control problem [16]–[18], [33]. In other words they base their designs on a robot model with velocity control inputs, rather than the more realistic model with torque control inputs. In doing so the controller is completely ignoring the vehicle dynamics due to mass, inertia and friction. This is known as the *perfect velocity tracking* assumption [19]. When it comes to the practical implementation of these kinematic controllers, this approach assumes that there is an independent low-level velocity control loop (usually implemented via a proportional-integral-derivative (PID) controller), that ascertains that the actual wheel velocities track precisely those requested by the kinematic control law [34]. However, while the use of independent PID velocity control loops is convenient and leads to acceptable performance in many applications involving slow-moving robots tracking low-acceleration trajectories, it can lead to high tracking errors, possibly resulting in total mission failure, in the face of more challenging tasks characterized by high reference velocities and accelerations [19]. In such situations, the robot nonlinear dynamics are no longer negligible and a better approach would be to replace the PID controller by a superior, though generally more complex, velocity controller whose design is based on a model relating the wheel velocities to the input torques. Such a controller would explicitly account for the vehicle's dynamic effects due to mass, friction and inertia. One such example is the well-established *computed-torque* approach [19], [34].

However, the dynamic model of a mobile robot is not only nonlinear but includes parameters or functions; such as mass, frictional terms and inertia; that are highly uncertain, time-varying or even unknown. Consequently, a number of adaptive control methods for the dynamic control of mobile robots have been proposed. These include both parametric adaptive control [20] and functional adaptive control [22], [25], [35]–[37]. The latter differs from the former in that the uncertainty is not restricted to parametric terms, but covers the dynamic functions themselves. We consider functional adaptive control to be more general and superior in handling higher degrees of uncertainty and unmodelled dynamics. Yet, all the mentioned adaptive robot controllers rely on the HCE assumption and so are prone to suffer from the aforementioned ill effects. In contrast, in our recent works [1], [26], we propose dual adaptive control techniques, rooted in computational intelligence, to address the problem of mobile robot control with uncertain/unknown dynamics.

In [26] we propose two novel dual control schemes employing two different kinds of artificial neural networks (ANNs), namely Gaussian radial basis functions (GaRBFs) and multilayer perceptrons (MLPs) [38], to estimate the WMR dynamic functions in real-time. The advantage of GaRBFs over MLPs lies in the fact that with GaRBFs the unknown ANN weights appear linearly in the stochastic state-space model formulated for estimation. This permits the use of the Kalman filter (KF) [39] for the recursive optimal ANN weight-tuning. However in the MLP case, this desirable property of linearity in the network parameters is not preserved, and the KF weight-tuning algorithm has to be replaced by a suboptimal nonlinear stochastic estimator, such as the extended Kalman filter (EKF) [40], which not only complicates the derivation of the control law, but introduces several approximations. On the other hand, unlike the activation functions employed in GaRBF ANNs, the sigmoidal functions in MLPs do not have localized receptive fields. This implies that typically MLP networks require less neurons than GaRBF ANNs to achieve the same degree of accuracy. Consequently, MLPs tend to be less computationally demanding, especially in the case of high-order systems, since the number of neurons need not rise exponentially with the number of states as in the case of GaRBF ANNs. The latter effect is known as the *curse of dimensionality* [41].

In the light of these arguments, the MLP dual adaptive scheme we proposed in [26] uses the EKF to estimate the nonlinearly-appearing ANN optimal parameters in real-time. The EKF approximates the state (in this case parameter) distribution by a Gaussian random variable (GRV) and propagates it analytically through the first-order linearization of

the nonlinear stochastic model. Moreover, the dual adaptive control law proposed for that scheme, is based on another first-order Taylor approximation of the measurement equation in the stochastic model. This adds further to the suboptimality of the proposed approach.

To lessen the extent of these approximations, in this paper, which extends on our recent preliminary work [1], we propose a novel MLP dual adaptive control scheme that uses a specifically devised form of the unscented Kalman filter (UKF) [42], [43] as a recursive weight-tuning algorithm, instead of the EKF employed in [26]. In addition, we propose a new dual adaptive control law that employs the unscented transform (UT) [42] to improve on the first-order Taylor approximation used in deriving the EKF-based controller in [26].

It should be pointed out that the convergence and stability analysis of dual adaptive control schemes presents a very difficult challenge, mainly due to the stochastic and adaptive nature of the problem. The few works that address these issues consider only linear systems of a particular form and are characterized by a number of nontrivial assumptions [9], [44]. Consequently, in contrast to the case of deterministic approaches, to prove convergence and stability for a dual adaptive nonlinear controller, is still considered to be an open problem within the research community. Hence in practice, as argued in [9], the stability of dual adaptive controllers is commonly demonstrated by computer simulations and real-life experiments.

The contribution of this paper comprises a detailed treatment of the two dual adaptive MLP control schemes mentioned previously and a set of verifying and comparative results, comprising realistic Mont Carlo simulations backed by rigourous statistical analysis and real-life experiments. In particular, we show that the proposed UT-based dual adaptive controller brings about significant improvements in tracking performance over the EKF-based dual adaptive scheme recently proposed in [26], while still employing the same computationally-friendly MLP architecture. To the best of our knowledge this is the first work in which the UT is being used in the context of dual adaptive control. In addition, one should note that very few adaptive controllers have ever been implemented and tested on a physical WMR, amongst which one finds [45], [46]. However, none of these address fully the uncertainty in the WMR dynamic functions nor take a dual adaptive control approach.

The rest of the paper is organized as follows. Section II contains preliminary material, including the development of the discrete-time dynamic model of the WMR considered in this work, and a formulation of the WMR trajectory tracking control problem. Section III presents the design of both the EKF-based and the proposed UT-based dual adaptive MLP control schemes. Monte Carlo simulation results supported by statistical hypothesis comparative tests and real-life experiments are then presented in Section IV, which is followed by a brief conclusion in Section V.

## II. PRELIMINARIES

In this work we address the trajectory tracking problem of the differentially driven WMR depicted in Figure 1. However, the framework we adopt in our design is completely modular. Consequently, the dual adaptive dynamic control scheme proposed in this paper can be easily adopted to address different navigation problems such as posture stabilization and path following [34], possibly even for different types of robotic configurations. In this section we outline the development of the dynamic model of the differentially driven WMR and formulate the trajectory tracking problem considered in this work.

### A. Modelling

With reference to the WMR configuration in Figure 1, we ignore the passive caster wheels and adopt the following notation throughout the article:

- $\mathrm{P_o}$: axle midpoint between the two wheels
- $\mathrm{P_c}$: centre of mass of the platform without wheels
- $d$: distance between $\mathrm{P_o}$ to $\mathrm{P_c}$
- $b$: distance from the centre of each wheel to $\mathrm{P_o}$
- $r$: radius of each wheel
- $m_c$: mass of the platform without wheels
- $m_w$: mass of each wheel
- $I_c$: moment of inertia of the platform about $\mathrm{P_c}$
- $I_w$: moment of inertia of each wheel about the axle
- $I_m$: moment of inertia of each wheel about its diameter

The robot coordinate vector is denoted by $\boldsymbol{q} = [x \;\; y \;\; \phi \;\; \theta_r \;\; \theta_l]^T$, where $(x, y)$ is the Cartesian coordinate of $\mathrm{P_o}$, $\phi$ is the robot's orientation with reference to the $x$-axis, and $\theta_r$, $\theta_l$ are the angular displacements about the axle of the right and left motorized wheels respectively. The *pose* of the robot refers to the vector $\boldsymbol{p} = [x \;\; y \;\; \phi]$.

*1) Kinematic Model:* The differential configuration of this WMR is subject to three kinematic constraints, stemming from

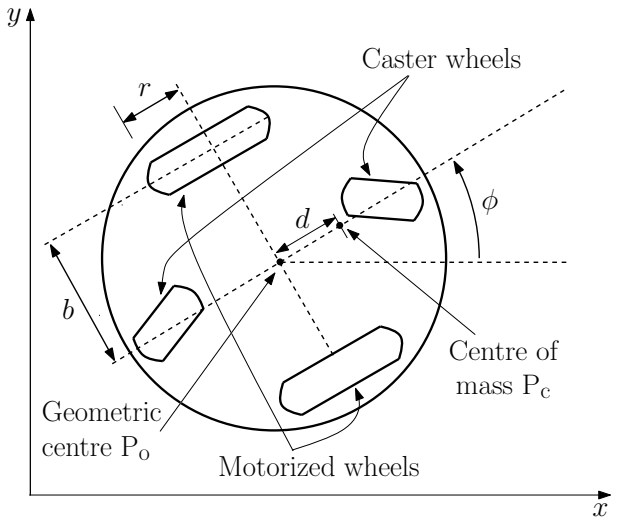



Fig. 1. Differentially driven wheeled mobile robot.

the fact that the translational velocity of the geometric centre $\mathrm{P_o}$ is always in the direction perpendicular to the driving axle, and the two driving wheels roll without slipping. The former leads to a holonomic constraint while the latter leads to two nonholonomic constrains [47]. Mathematically this is described by $\boldsymbol{A}(\boldsymbol{q})\dot{\boldsymbol{q}} = \boldsymbol{0}$, where

$$\boldsymbol{A}(\boldsymbol{q}) = \begin{bmatrix} -\sin\phi & \cos\phi & 0 & 0 & 0 \\ \cos\phi & \sin\phi & b & -r & 0 \\ \cos\phi & \sin\phi & -b & 0 & -r \end{bmatrix}.$$

These three kinematic constraints, along with a few other relationships arising from the geometry of the WMR depicted in Figure 1, can be used to show that the kinematic model of this differentially driven WMR is given by

$$\dot{\boldsymbol{q}} = \boldsymbol{S}(\boldsymbol{q})\boldsymbol{\nu}, \tag{1}$$

where

$$\boldsymbol{S}(\boldsymbol{q}) = \begin{bmatrix} \frac{r}{2}\cos\phi & \frac{r}{2}\cos\phi \\ \frac{r}{2}\sin\phi & \frac{r}{2}\sin\phi \\ \frac{r}{2b} & -\frac{r}{2b} \\ 1 & 0 \\ 0 & 1 \end{bmatrix},$$

and $\boldsymbol{\nu}$ denotes a vector composed of the angular velocities of the two motorized wheels, that is, $\boldsymbol{\nu} = \begin{bmatrix} \nu_r & \nu_l \end{bmatrix}^T = \begin{bmatrix} \dot{\theta}_r & \dot{\theta}_l \end{bmatrix}^T$. It is important to note that:

**Remark II.1.** *The two independent columns of $\boldsymbol{S}(\boldsymbol{q})$ are in the null space of $\boldsymbol{A}(\boldsymbol{q})$, that is, $\boldsymbol{A}(\boldsymbol{q})\boldsymbol{S}(\boldsymbol{q}) = \boldsymbol{0}$.*

*2) Dynamic Model:* The equations of motion of this WMR can be derived using Lagrangian mechanics. The Euler-Lagrange equation for the nonholonomic WMR considered in this paper is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial K}{\partial \dot{q}_i}\right) - \frac{\partial K}{\partial q_i} = Q_i - \sum_{c=1}^{3} a_{ci}\lambda_c, \quad (i = 1, 2, \ldots, 5), \tag{2}$$

where $K(\boldsymbol{q}, \dot{\boldsymbol{q}})$ is the total kinetic energy of the WMR, $q_i$ is the $i^{\text{th}}$ element of the coordinate vector $\boldsymbol{q}$, $Q_i$ is the $i^{\text{th}}$ Lagrange force, $a_{ci}$ is the $(c,i)^{\text{th}}$ element of the constraints matrix $\boldsymbol{A}(\boldsymbol{q})$ and $\lambda_c$ is the $c^{\text{th}}$ element of the vector of Lagrange multipliers $\boldsymbol{\lambda}$. It can be shown that the total kinetic energy of the WMR is given by

$$\begin{aligned} K(\boldsymbol{q}, \dot{\boldsymbol{q}}) &= \frac{m}{2}\left(\dot{x}^2 + \dot{y}^2\right) + m_c d\dot{\phi}\left(\dot{y}\cos\phi - \dot{x}\sin\phi\right) \\ &\quad + \frac{I}{2}\dot{\phi}^2 + \frac{I_w}{2}\left(\dot{\theta_r}^2 + \dot{\theta_l}^2\right), \end{aligned} \tag{3}$$

where $m = m_c + 2m_w$, $I = (I_c + m_c d^2) + 2(I_m + m_w b^2)$. Equation (3) can then be used to work out the derivative terms in (2). This leads to the equations of motion of the WMR, given by:

$$\boldsymbol{M}(\boldsymbol{q})\ddot{\boldsymbol{q}} + \boldsymbol{V}(\boldsymbol{q}, \dot{\boldsymbol{q}}) = \boldsymbol{E}\boldsymbol{\tau} - \boldsymbol{A}^T(\boldsymbol{q})\boldsymbol{\lambda}, \tag{4}$$

where:

$$\boldsymbol{M}(\boldsymbol{q}) = \begin{bmatrix} m & 0 & -m_c d\sin\phi & 0 & 0 \\ 0 & m & m_c d\cos\phi & 0 & 0 \\ -m_c d\sin\phi & m_c d\cos\phi & I & 0 & 0 \\ 0 & 0 & 0 & I_w & 0 \\ 0 & 0 & 0 & 0 & I_w \end{bmatrix},$$

$$\boldsymbol{V}(\boldsymbol{q}, \dot{\boldsymbol{q}}) = \begin{bmatrix} -m_c d\dot{\phi}^2\cos\phi \\ -m_c d\dot{\phi}^2\sin\phi \\ 0 \\ 0 \\ 0 \end{bmatrix}, \boldsymbol{E} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and $\boldsymbol{\tau} = \begin{bmatrix} \tau_r & \tau_l \end{bmatrix}^T$ is the torque vector, with $\tau_r$ and $\tau_l$ denoting the torques applied to the right and left wheels respectively.

The kinematic model in (1) and the equations of motion in (4) can be used to determine the WMR dynamics relating the wheels acceleration $\dot{\boldsymbol{\nu}}$ to the wheels torque $\boldsymbol{\tau}$ as follows [47]. We differentiate (1) with respect to time, substitute the expression for $\ddot{\boldsymbol{q}}$ in (4), and finally pre-multiply the resulting expression by $\boldsymbol{S}^T$. Noting that: $\boldsymbol{S}^T\boldsymbol{A}^T = \boldsymbol{0}$ (by Remark II.1), $\boldsymbol{S}^T\boldsymbol{E} = \mathbf{I_2}$ (where throughout the paper $\mathbf{I_i}$ denotes an $(i \times i)$ identity matrix), and $\dot{\phi} = \frac{r}{2b}(\nu_r - \nu_l)$ (by (1)); the resulting dynamic model can be expressed by

$$\bar{\boldsymbol{M}}\dot{\boldsymbol{\nu}} + \bar{\boldsymbol{V}}(\boldsymbol{\nu}) + \bar{\boldsymbol{F}}(\boldsymbol{\nu}) = \boldsymbol{\tau}, \tag{5}$$

where:

$$\bar{\boldsymbol{M}} = \boldsymbol{S^T M S} = \begin{bmatrix} \frac{r^2}{4b^2}(mb^2 + I) + I_w & \frac{r^2}{4b^2}(mb^2 - I) \\ \frac{r^2}{4b^2}(mb^2 - I) & \frac{r^2}{4b^2}(mb^2 + I) + I_w \end{bmatrix},$$

$$\bar{\boldsymbol{V}}(\boldsymbol{\nu}) = \boldsymbol{S^T M}\dot{\boldsymbol{S}}\boldsymbol{\nu} + \boldsymbol{S^T V} = \frac{m_c d r^3}{4b^2}\begin{bmatrix} \nu_r\nu_l - \nu_l^2 \\ \nu_r\nu_l - \nu_r^2 \end{bmatrix},$$

and $\bar{\boldsymbol{F}}(\boldsymbol{\nu})$ is introduced to account for any wheel velocity-dependent frictional terms.

**Remark II.2.** *$\bar{\boldsymbol{M}}$ is symmetric, positive definite, and is independent of the coordinate vector and/or its derivatives.*

**Remark II.3.** *In general, $\bar{\boldsymbol{V}}(\boldsymbol{\nu})$ and $\bar{\boldsymbol{F}}(\boldsymbol{\nu})$ are the two terms that render the WMR dynamics nonlinear.*

To account for the fact that the controller is to be implemented on a digital computer, the continuous-time dynamics (5) are discretized through a first-order forward Euler approximation with a sampling interval of $T$ seconds. The resulting nonlinear discrete-time dynamic model is given by

$$\boldsymbol{\nu}_k - \boldsymbol{\nu}_{k-1} = \boldsymbol{f}_{k-1} + \boldsymbol{G}_{k-1}\boldsymbol{\tau}_{k-1}, \tag{6}$$

where the subscript integer $k$ denotes that the corresponding variable is evaluated at time $kT$ seconds, and vector $\boldsymbol{f}_{k-1}$ and matrix $\boldsymbol{G}_{k-1}$, which together encapsulate the WMR dynamics, are given by

$$\begin{aligned} \boldsymbol{f}_{k-1} &= -T\bar{\boldsymbol{M}}_{k-1}^{-1}\left(\bar{\boldsymbol{V}}_{k-1} + \bar{\boldsymbol{F}}_{k-1}\right), \\ \boldsymbol{G}_{k-1} &= T\bar{\boldsymbol{M}}_{k-1}^{-1}. \end{aligned} \tag{7}$$

The following conditions are assumed to hold:

**Assumption II.1.** *The control input vector $\boldsymbol{\tau}$ remains constant over a sampling interval of $T$ seconds (zero-order hold).*

**Assumption II.2.** *The sampling interval $T$ is chosen low enough for the Euler approximation error to be negligible.*

### B. Trajectory Tracking

The trajectory tracking task of WMRs is commonly defined via the concept of the *virtual vehicle* [17]. In this formulation, the time-dependent reference trajectory is designated by a nonstationary virtual vehicle, *kinematically identical* to the real vehicle. The control task is for the real vehicle to track the virtual vehicle at all times, *in both pose and velocity*. It is important to note that this problem is different and generally more challenging than path-following. This stems from the fact that in trajectory tracking the reference path is time-indexed (hence dictating speed as well as position), while in path-following the reference contains no temporal information and the vehicle speed is typically fixed and predetermined [34].

The trajectory tracking error in discrete-time is commonly defined by a tracking error vector $\boldsymbol{e}_k = [e_{1k}\ e_{2k}\ e_{3k}]^T$, expressed pictorially in Figure 2, and mathematically defined by

$$\boldsymbol{e}_k = \begin{bmatrix} \cos\phi_k & \sin\phi_k & 0 \\ -\sin\phi_k & \cos\phi_k & 0 \\ 0 & 0 & 1 \end{bmatrix} (\boldsymbol{p_r}_k - \boldsymbol{p}_k), \qquad (8)$$

where $\boldsymbol{p_r}_k = [x_{rk}\ \ y_{rk}\ \ \phi_{rk}]^T$ denotes the virtual vehicle pose vector. Hence, in trajectory tracking the objective is to make $\boldsymbol{e}_k$ converge to zero, so that $\boldsymbol{p}_k$ converges to $\boldsymbol{p_r}_k$.

## III. Control Design

As argued in Section I, the motion control of WMRs is commonly addressed as two separate tasks, namely kinematic and dynamic control [19], [34], [37]. Kinematic control is concerned solely with the steering system (1). Specifically its aim is to devise a control law for the robot wheel velocities, so as to stabilize the pose of the robot as required by the navigation task at hand; be it trajectory tracking, path-following or posture stabilization. In the case of trajectory tracking, the aim of the kinematic controller is to compute the wheel velocities required to minimize the robot tracking error $\boldsymbol{e}_k$. On the other hand, the aim of the dynamic controller is to compute the wheel torques required in order to ensure that the robot accurately tracks the velocities computed by the kinematic controller. Hence, the two control loops operate in cascade; with the kinematic controller's output (a velocity command) serving as the reference input of the cascaded dynamic controller, which computes the torque required to drive the WMR wheels at the specified velocities. This approach renders the overall control architecture modular, since the kinematic controller, which is specific to the navigation problem at hand, can be easily replaced while still retaining the same dynamic controller. In our work we adopt this modular architecture, depicted in Figure 3, and design the dynamic controller to be dual adaptive as detailed in the rest of this section.

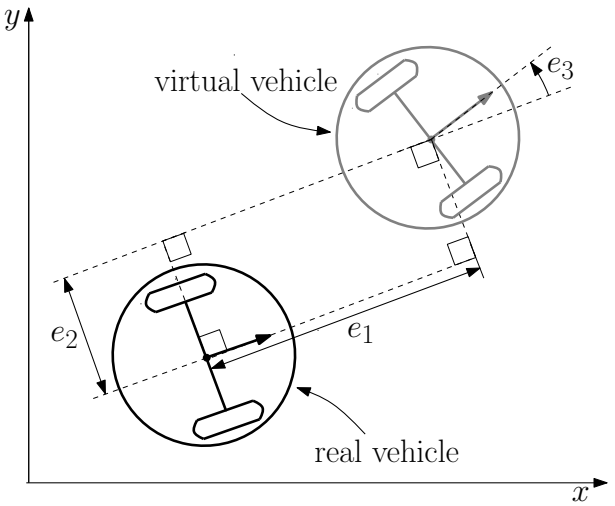



Fig. 2. Trajectory tracking via the concept of the virtual vehicle.

### A. The Kinematic Controller

As argued earlier, the role of the kinematic controller in trajectory tracking is to make $\boldsymbol{e}_k$ converge to zero, so that $\boldsymbol{p}_k$ converges to $\boldsymbol{p_r}_k$. To address this well-researched problem we opt to adopt an established kinematic controller, originally presented in [17], and convert it to discreet-time so as to integrate it in our formulation. The resulting kinematic control law is given by

$$\boldsymbol{\nu_c}_k = \boldsymbol{C} \begin{bmatrix} v_{rk}\cos e_{3k} + k_1 e_{1k} \\ \omega_{rk} + k_2 v_{rk} e_{2k} + k_3 v_{rk} \sin e_{3k} \end{bmatrix}, \qquad (9)$$

where $\boldsymbol{\nu_c}_k$ is the wheel velocity command vector issued by the kinematic controller, $k_1$, $k_2$, and $k_3$ are positive design parameters, $v_{rk} > 0$ and $\omega_{rk}$ are the translational and angular *virtual vehicle* velocities respectively (assumed to be *continuous* functions, at least know one sampling interval ahead), and $\boldsymbol{C}$ is a velocity conversion matrix given by

$$\boldsymbol{C} = \begin{bmatrix} \frac{1}{r} & \frac{b}{r} \\ \frac{1}{r} & -\frac{b}{r} \end{bmatrix}.$$

Stability analysis and the corresponding necessary conditions of this controller in continuous-time are detailed in [17].

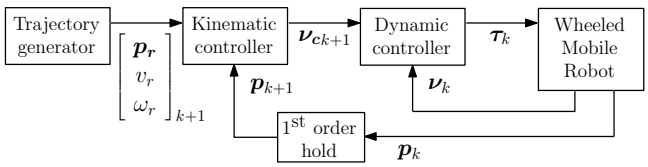



Fig. 3. Dynamic control architecture.

### B. Nonadaptive Dynamic Control

If the nonlinear dynamic functions $\boldsymbol{f}_k$ and $\boldsymbol{G}_k$ are perfectly known, the computed-torque control law

$$\boldsymbol{\tau}_k = \boldsymbol{G}_k^{-1}\left(\boldsymbol{\nu}_{\boldsymbol{c}k+1} - \boldsymbol{\nu}_k - \boldsymbol{f}_k + k_d\left(\boldsymbol{\nu}_{\boldsymbol{c}k} - \boldsymbol{\nu}_k\right)\right), \qquad (10)$$

with the design parameter $-1 < k_d < 1$, yields the following closed-loop stable linear dynamics

$$\boldsymbol{\nu}_{k+1} = \boldsymbol{\nu}_{\boldsymbol{c}k+1} + k_d\left(\boldsymbol{\nu}_{\boldsymbol{c}k} - \boldsymbol{\nu}_k\right),$$

when substituted in the dynamic model in (6). This ensures that $|\boldsymbol{\nu}_{\boldsymbol{c}k} - \boldsymbol{\nu}_k| \to 0$ as $k \to \infty$. It is important to note that:

**Remark III.1.** *Control law (10) requires the velocity command vector $\boldsymbol{\nu}_{\boldsymbol{c}k+1}$ to be available at instant $k$. For this reason, the kinematic control law (9) is advanced by one sampling interval. This means that at instant $k$, the values of $v_{rk+1}$, $\omega_{rk+1}$ and $\boldsymbol{e}_{k+1}$ need to be known. Additionally, from (8) it is clear that $\boldsymbol{p}_{\boldsymbol{r}k+1}$ and $\boldsymbol{p}_{k+1}$ are needed to determine $\boldsymbol{e}_{k+1}$. Having the values of reference signal $\boldsymbol{p}_{\boldsymbol{r}k+1}$, $v_{rk+1}$ and $\omega_{rk+1}$ available at instant $k$ is easy, since it simply means that the path-planning algorithm is required to generate the reference trajectory one sampling interval ahead. On the other hand, for the non-reference signal $\boldsymbol{p}_{k+1}$, we propose to estimate its value via the first-order approximation $\boldsymbol{p}_{k+1} \approx 2\boldsymbol{p}_k - \boldsymbol{p}_{k-1}$. This is justified in the light of Assumption II.2.*

**Remark III.2.** *The case with $k_d = 0$ in (10), corresponds to* deadbeat control *associated with digital control systems [48].*

### C. Dual Adaptive Dynamic Control using MLPs

The computed-torque dynamic control law (10) driven by the kinematic law (9), is a solution to the trajectory tracking problem *only if* the WMR dynamic functions $\boldsymbol{f}_{k-1}$ and $\boldsymbol{G}_{k-1}$ in (6) are perfectly known. As emphasized in Section I, this is rarely the case in real-life robotic applications commonly exhibiting: unmodelled dynamics, unknown/time-varying parameters, and imperfect/noisy sensor measurements. Most works, address these issues via some form of HCE adaptive control. In contrast, the two schemes presented in this paper not only consider $\boldsymbol{f}_{k-1}$ and $\boldsymbol{G}_{k-1}$ to be completely unknown to the controller, but also feature dual adaptive properties to handle the issue of uncertainty as explained in Section I. The two dual adaptive schemes, detailed in this section, both employ a stochastically-trained ANN-based algorithm to approximate these functions recursively in real-time.

Specifically, a sigmoidal MLP ANN with one hidden layer is used to approximate the nonlinear vector-valued function $\boldsymbol{f}_{k-1}$, as depicted in Figure 4. Its output is given by

$$\tilde{\boldsymbol{f}}_{k-1} = \begin{bmatrix} \boldsymbol{\phi}^T(\boldsymbol{x}_{k-1}, \hat{\boldsymbol{a}}_k)\hat{\boldsymbol{w}}_{\boldsymbol{1}k} \\ \boldsymbol{\phi}^T(\boldsymbol{x}_{k-1}, \hat{\boldsymbol{a}}_k)\hat{\boldsymbol{w}}_{\boldsymbol{2}k} \end{bmatrix}, \qquad (11)$$

in the light of the following statements:

**Definition III.1.** *$\boldsymbol{x}_{k-1} = [\boldsymbol{\nu}_{k-1} \quad 1]$ denotes the ANN input. The augmented constant serves as a bias input. This selection of the ANN input stems from the fact that $\boldsymbol{f}_{k-1}$ is effectively a function of $\boldsymbol{\nu}_{k-1}$.*

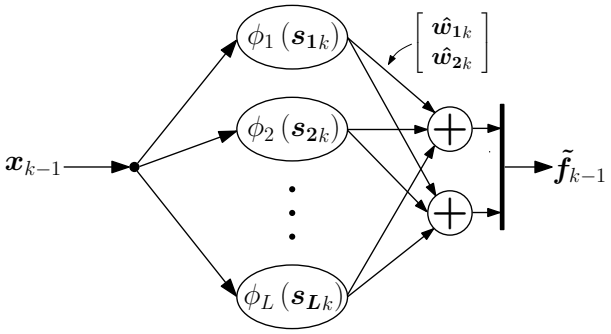



Fig. 4. Sigmoidal Multilayer Perceptron neural network.

**Definition III.2.** *$\boldsymbol{\phi}(\cdot,\cdot)$ is the vector of sigmoidal activation functions, whose $i^{th}$ element is given by $\phi_i = 1/\left(1 + \exp\left(-\hat{\boldsymbol{s}}_{\boldsymbol{i}}^T\boldsymbol{x}\right)\right)$, where $\hat{\boldsymbol{s}}_{\boldsymbol{i}}$ is $i^{th}$ vector element in the group vector $\hat{\boldsymbol{a}}$;* i.e. *$\hat{\boldsymbol{a}} = \left[\hat{\boldsymbol{s}}_{\boldsymbol{1}}^T \quad \cdots \quad \hat{\boldsymbol{s}}_{\boldsymbol{L}}^T\right]^T$ where $L$ denotes the number of neurons. The time index has been dropped for clarity, and throughout the paper the $\hat{}$ notation indicates that the operand is undergoing estimation.*

**Definition III.3.** *$\hat{\boldsymbol{w}}_{\boldsymbol{i}k}$ represents the synaptic weight estimate vector of the connection between the neuron hidden layer and the $i^{th}$ output element of the ANN.*

**Assumption III.1.** *The input vector $\boldsymbol{x}_{k-1}$ is contained within a known, arbitrarily large compact set $\boldsymbol{\chi} \subset \mathbb{R}^2$. This is justified since the wheel velocities are inherently bounded.*

Moreover, it is known that $\boldsymbol{G}_{k-1}$ is a state-independent matrix with unknown elements (refer to (7)). Hence, its estimation does not require the use of an ANN. In addition it is a symmetric matrix, a property which is exploited to construct its estimate as follows

$$\tilde{\boldsymbol{G}}_{k-1} = \begin{bmatrix} \hat{g}_{1k-1} & \hat{g}_{2k-1} \\ \hat{g}_{2k-1} & \hat{g}_{1k-1} \end{bmatrix}, \qquad (12)$$

where $\hat{g}_{1k-1}$ and $\hat{g}_{2k-1}$ represent the estimates of the unknown elements in $\boldsymbol{G}_{k-1}$.

We formulate the ANN weight-tuning task as a stochastic nonlinear estimation problem. The following preliminaries are necessary in order to proceed.

**Definition III.4.** *The unknown parameters requiring estimation are grouped in a single vector $\hat{\boldsymbol{z}}_k = \left[\hat{\boldsymbol{r}}_k^T \quad \hat{\boldsymbol{g}}_k^T\right]^T$, where $\hat{\boldsymbol{r}}_k = \left[\hat{\boldsymbol{w}}_{\boldsymbol{1}k}^T \quad \hat{\boldsymbol{w}}_{\boldsymbol{2}k}^T \quad \hat{\boldsymbol{a}}_k^T\right]^T$ and $\hat{\boldsymbol{g}}_k = \left[\hat{g}_{1k-1} \quad \hat{g}_{2k-1}\right]^T$.*

**Definition III.5.** *The measured output in the dynamic model (6) is denoted by $\boldsymbol{y}_k = \boldsymbol{\nu}_k - \boldsymbol{\nu}_{k-1}$. In our practical implementation $\boldsymbol{\nu}_k$ is acquired from the wheel encoders.*

**Assumption III.2.** *By the* Universal Approximation Theorem *of ANN, inside the compact set $\boldsymbol{\chi}$, the ANN approximation error is negligibly small when the estimate $\hat{\boldsymbol{r}}_k$ is equal to some unknown optimal vector $\boldsymbol{r}_k^*$. The $^*$ notation denotes optimality.*

In view of the stochastic adaptive approach taken in this work, the unknown optimal parameter vector $\boldsymbol{z}_k^*$ is

treated as a random variable, with the initial condition $p(\boldsymbol{z}_0^*) \sim \mathcal{N}(\hat{\boldsymbol{z}}_0, \boldsymbol{P}_0)$, meaning that $\boldsymbol{z}_0^*$ is normally distributed with mean $\hat{\boldsymbol{z}}_0$ and covariance $\boldsymbol{P}_0$. This notation is adopted throughout the article. Effectively, the covariance value $\boldsymbol{P}_0$ reflects the confidence in the initial guess $\hat{\boldsymbol{z}}_0$.

By (11), (12), all previous definitions and assumptions, it follows that the model in (6) can be represented in the following stochastic state-space form

$$\begin{aligned} \boldsymbol{z}_{k+1}^* &= \boldsymbol{z}_k^* + \boldsymbol{\rho}_k \\ \boldsymbol{y}_k &= \boldsymbol{h}\left(\boldsymbol{x}_{k-1}, \boldsymbol{\tau}_{k-1}, \boldsymbol{z}_k^*\right) + \boldsymbol{\epsilon}_k, \end{aligned} \tag{13}$$

where the vector-valued function $\boldsymbol{h}\left(\boldsymbol{x}_{k-1}, \boldsymbol{\tau}_{k-1}, \boldsymbol{z}_k^*\right)$ is nonlinear in $\boldsymbol{z}_k^*$, and is given by

$$\boldsymbol{h}\left(\cdot\right) = \tilde{\boldsymbol{f}}(\boldsymbol{x}_{k-1}, \boldsymbol{r}_k^*) + \tilde{\boldsymbol{G}}(\boldsymbol{g}_k^*)\boldsymbol{\tau}_{k-1}. \tag{14}$$

In this model, the unknown optimal parameter vector $\boldsymbol{z}_k^*$ is characterized as a stationary process corrupted by an artificial process noise $\boldsymbol{\rho}_k$, which aids convergence and tracking during estimation. In addition, observation uncertainty is catered for by augmenting a random measurement noise $\boldsymbol{\epsilon}_k$ to $\boldsymbol{y}_k$.

It is evident, from (14), that the use of the MLP ANN, which brings about certain practical advantages over GaRBF as argued in Section I, results in a *nonlinear* measurement equation in the stochastic state-space model (13) formulated for estimation. In order to address this issue in a stochastic framework, we have to employ a nonlinear recursive estimator.

The two dual adaptive schemes presented in this paper depart from this point in our formulation and proceed to tackle the estimation and control problems in different ways, as detailed next.

*1) EKF-based Dual Adaptive Scheme:* For the sake of clarity and completeness, the MLP dual adaptive scheme proposed in [26] and used for comparisons in this paper is revisited in this section. In this scheme, we employ the EKF in prediction mode for the recursive real-time estimation of $\boldsymbol{z}_{k+1}^*$ as follows.

**Definition III.6.** *The* information state *denoted by $I^k$, consists of all measurements up to instant $k$ and all previous inputs.*

**Assumption III.3.** *$\boldsymbol{\epsilon}_k$ and $\boldsymbol{\rho}_k$ are both zero-mean white Gaussian processes with covariances $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ and $\boldsymbol{Q}_{\boldsymbol{\rho}}$ respectively. Moreover $\boldsymbol{\epsilon}_k$, $\boldsymbol{\rho}_k$ and $\boldsymbol{z}_0^*$ are mutually independent $\forall k$.*

**Lemma III.1.** *In the light of (13), Definition III.6, and Assumption III.3, it follows that $p(\boldsymbol{z}_{k+1}^*|I^k) \approx \mathcal{N}(\hat{\boldsymbol{z}}_{k+1}, \boldsymbol{P}_{k+1})$, where $\hat{\boldsymbol{z}}_{k+1}$ and $\boldsymbol{P}_{k+1}$ are computed at each control step according to the EKF Algorithm III.1. Consequently, $\hat{\boldsymbol{z}}_{k+1}$ is considered to be the estimate of $\boldsymbol{z}_{k+1}^*$ conditioned on $I^k$, and $\boldsymbol{P}_{k+1}$ can be viewed as a measure of this estimate's uncertainty.*

*Proof:* The proof of this lemma follows directly that of the EKF in prediction mode, when applied to the nonlinear stochastic state-space model in (13). ■

Given the previous prediction $\left(\hat{\boldsymbol{z}}_{k|k-1}, \boldsymbol{P}_{k|k-1}\right)$; denoted in short-form by $(\hat{\boldsymbol{z}}_k, \boldsymbol{P}_k)$; the following EKF (prediction mode) algorithm generates the new prediction $(\hat{\boldsymbol{z}}_{k+1}, \boldsymbol{P}_{k+1})$.

*1) Evaluating $\boldsymbol{\nabla}_{\boldsymbol{h}k}$, the Jacobian matrix of $\boldsymbol{h}\left(\boldsymbol{x}_{k-1}, \boldsymbol{\tau}_{k-1}, \boldsymbol{z}_k^*\right)$ with respect to $\boldsymbol{z}_k^*$ evaluated at $\hat{\boldsymbol{z}}_k$:*

$$\boldsymbol{\nabla}_{\boldsymbol{h}k} \triangleq \left[\boldsymbol{\nabla}_{\boldsymbol{f}k} \;\; \boldsymbol{\nabla}_{\boldsymbol{\Gamma}k}\right] = \left[\left.\frac{\partial(\tilde{\boldsymbol{f}}_{k-1})}{\partial(\boldsymbol{r}_k^*)}\right|_{\hat{\boldsymbol{r}}_k} \;\; \left.\frac{\partial(\tilde{\boldsymbol{G}}_{k-1}\boldsymbol{\tau}_{k-1})}{\partial(\boldsymbol{g}_k^*)}\right|_{\hat{\boldsymbol{g}}_k}\right],$$

where by (11), (12) and (14), it can be shown that:

$$\begin{aligned} \boldsymbol{\nabla}_{\boldsymbol{f}k} &= \begin{bmatrix} \boldsymbol{\phi}_{k-1}^T & \boldsymbol{0}^T & \cdots \hat{w_{1,i}}(\phi_i - {\phi_i}^2)\boldsymbol{x}^T \cdots \\ \boldsymbol{0}^T & \boldsymbol{\phi}_{k-1}^T & \cdots \hat{w_{2,i}}(\phi_i - {\phi_i}^2)\boldsymbol{x}^T \cdots \end{bmatrix}, \\ \boldsymbol{\nabla}_{\boldsymbol{\Gamma}k} &= \begin{bmatrix} \tau_{r\,k-1} & \tau_{l\,k-1} \\ \tau_{l\,k-1} & \tau_{r\,k-1} \end{bmatrix}, \end{aligned} \tag{15}$$

where: $i = 1, \ \ldots \ , L$, $\hat{w_{j,i}}$ denotes the $i^{\text{th}}$ element of the $j^{\text{th}}$ output weight vector $\hat{\boldsymbol{w}}_{\boldsymbol{j}k}$, notation-wise $\boldsymbol{\phi}_{k-1}$ implies that the activation function is evaluated for $\boldsymbol{x}_{k-1}$ and $\hat{\boldsymbol{a}}_k$, $\boldsymbol{0}$ denotes a zero-vector of the same length as $\boldsymbol{\phi}_{k-1}$, and $\phi_i$ and $\boldsymbol{x}$ both correspond to time instant $(k-1)$.

*2) Performing the prediction step:*

$$\begin{aligned} \hat{\boldsymbol{z}}_{k+1} &= \hat{\boldsymbol{z}}_k + \boldsymbol{K}_k \boldsymbol{i}_k \\ \boldsymbol{P}_{k+1} &= \boldsymbol{P}_k - \boldsymbol{K}_k \boldsymbol{\nabla}_{\boldsymbol{h}k} \boldsymbol{P}_k + \boldsymbol{Q}_{\boldsymbol{\rho}} \end{aligned} \tag{16}$$

where the Kalman gain and the innovation vector are respectively given by:

$$\begin{aligned} \boldsymbol{K}_k &= \boldsymbol{P}_k \boldsymbol{\nabla}_{\boldsymbol{h}k}^T \left(\boldsymbol{\nabla}_{\boldsymbol{h}k} \boldsymbol{P}_k \boldsymbol{\nabla}_{\boldsymbol{h}k}^T + \boldsymbol{R}_{\boldsymbol{\epsilon}}\right)^{-1} \\ \boldsymbol{i}_k &= \boldsymbol{y}_k - \boldsymbol{h}\left(\boldsymbol{x}_{k-1}, \boldsymbol{\tau}_{k-1}, \hat{\boldsymbol{z}}_k\right). \end{aligned}$$

**Algorithm III.1:** The EKF parameter-prediction algorithm.

**Lemma III.2.** *On the basis of Lemma III.1, it follows that $p(\boldsymbol{y}_{k+1}|I^k)$ is approximately Gaussian with mean $\boldsymbol{h}\left(\boldsymbol{x}_k, \boldsymbol{\tau}_k, \hat{\boldsymbol{z}}_{k+1}\right)$ and covariance $\boldsymbol{\nabla}_{\boldsymbol{h}k+1} \boldsymbol{P}_{k+1} \boldsymbol{\nabla}_{\boldsymbol{h}k+1}^T + \boldsymbol{R}_{\boldsymbol{\epsilon}}$.*

*Proof:* Expressing $\boldsymbol{y}_{k+1}$ as a first-order Taylor series around $\boldsymbol{z}_{k+1}^* = \hat{\boldsymbol{z}}_{k+1}$ yields the following approximation

$$\boldsymbol{y}_{k+1} \approx \boldsymbol{h}\left(\boldsymbol{x}_k, \boldsymbol{\tau}_k, \hat{\boldsymbol{z}}_{k+1}\right) + \boldsymbol{\nabla}_{\boldsymbol{h}k+1}\left(\boldsymbol{z}_{k+1}^* - \hat{\boldsymbol{z}}_{k+1}\right) + \boldsymbol{\epsilon}_{k+1}.$$

Noting that $\boldsymbol{z}_{k+1}^*$ and $\boldsymbol{\epsilon}_{k+1}$ are the only probabilistic terms on the right-hand side of this approximation, the expected value of $\boldsymbol{y}_{k+1}$ conditioned on $I^k$, denoted by $E\left\{\boldsymbol{y}_{k+1}|I^k\right\}$, can be expressed as a sum of three terms:

$$\boldsymbol{h}\left(\boldsymbol{x}_k, \boldsymbol{\tau}_k, \hat{\boldsymbol{z}}_{k+1}\right) + \boldsymbol{\nabla}_{\boldsymbol{h}k+1}\left(E\left\{\boldsymbol{z}_{k+1}^*\right\} - \hat{\boldsymbol{z}}_{k+1}\right) + E\left\{\boldsymbol{\epsilon}_{k+1}\right\}.$$

Since $E\left\{\boldsymbol{z}_{k+1}^*\right\} = \hat{\boldsymbol{z}}_{k+1}$, by Lemma III.1, and $E\left\{\boldsymbol{\epsilon}_{k+1}\right\} = 0$, by Assumption III.3, the second and third term are both equal to zero, leaving the first term as the mean value of $p(\boldsymbol{y}_{k+1}|I^k)$. Using the same Taylor series approximation, we note that the covariance of the right-hand side can be written as

$\text{Cov}\left(\boldsymbol{\nabla}_{\boldsymbol{h}_{k+1}}\boldsymbol{z}^*_{k+1}\right) + \text{Cov}\left(\boldsymbol{\epsilon}_{k+1}\right)$ which by Lemma III.1 and Assumption III.3 reduces to $\boldsymbol{\nabla}_{\boldsymbol{h}_{k+1}}\boldsymbol{P}_{k+1}\boldsymbol{\nabla}^T_{\boldsymbol{h}_{k+1}} + \boldsymbol{R}_{\boldsymbol{\epsilon}}$. ■

Algorithm III.1, in view of Lemma III.1, constitutes the adaptation law for the EKF-based dual adaptive scheme. Moreover, by Lemma III.2, this algorithm provides a real-time update of the probability density function $p(\boldsymbol{y}_{k+1}|I^k)$, which is used to develop the dual adaptive control law as follows.

The explicit-type suboptimal innovation-based performance index $J_{inn}$, adopted from [8], and modified to fit our multiple-input multiple-output (MIMO) nonlinear problem, is given by

$$\begin{aligned} J_{inn} &= E\Big\{ \left(\boldsymbol{y}_{k+1} - \boldsymbol{y}_{\boldsymbol{d}_{k+1}}\right)^T \boldsymbol{Q_1} \left(\boldsymbol{y}_{k+1} - \boldsymbol{y}_{\boldsymbol{d}_{k+1}}\right) \\ &\quad + \left(\boldsymbol{\tau}_k^T \boldsymbol{Q_2} \boldsymbol{\tau}_k\right) + \left(\boldsymbol{i}_{k+1}^T \boldsymbol{Q_3} \boldsymbol{i}_{k+1}\right) \Big| I^k \Big\}, \end{aligned} \tag{17}$$

in view of the following definitions:

**Definition III.7.** *$\boldsymbol{y}_{\boldsymbol{d}_{k+1}}$ is the reference vector of $\boldsymbol{y}_{k+1}$ and is given by $\boldsymbol{y}_{\boldsymbol{d}_{k+1}} = \boldsymbol{\nu}_{\boldsymbol{c}k+1} - \boldsymbol{\nu}_k$.*

**Definition III.8.** *Design parameters $\boldsymbol{Q_1}$, $\boldsymbol{Q_2}$ and $\boldsymbol{Q_3}$ are diagonal and $\in \mathbb{R}^{2\times 2}$. Additionally: $\boldsymbol{Q_1}$ is positive definite, $\boldsymbol{Q_2}$ is positive semi-definite, and $-\boldsymbol{Q_1} \leq \boldsymbol{Q_3} \leq \boldsymbol{0}$ (element-wise).*

**Remark III.3.** *The design parameter $\boldsymbol{Q_1}$ is introduced to penalize tracking errors, $\boldsymbol{Q_2}$ induces a penalty on large control inputs, and $\boldsymbol{Q_3}$ affects the innovation vector so as to induce the dual adaptive feature characterizing this stochastic control law.*

The EKF-based dual adaptive control law is given by:

**Theorem III.1.** *The control law minimizing performance index $J_{inn}$ in (17), subject to the WMR dynamic model (5) and all the previous definitions, assumptions and lemmas in this formulation, is given by*

$$\begin{aligned} \boldsymbol{\tau}_k &= \left(\tilde{\boldsymbol{G}}_k^T \boldsymbol{Q_1} \tilde{\boldsymbol{G}}_k + \boldsymbol{Q_2} + \boldsymbol{N}_{k+1}\right)^{-1} \\ &\quad \times \left(\tilde{\boldsymbol{G}}_k^T \boldsymbol{Q_1} \left(\boldsymbol{y}_{\boldsymbol{d}_{k+1}} - \tilde{\boldsymbol{f}}_k\right) - \boldsymbol{\kappa}_{k+1}\right), \end{aligned} \tag{18}$$

*where $\tilde{\boldsymbol{f}}_k$ and $\tilde{\boldsymbol{G}}_k$ are computed via (11) and (12) using the latest estimate vector $\hat{\boldsymbol{z}}_{k+1}$ given by Algorithm III.1, and $\boldsymbol{\kappa}_{k+1}$ and $\boldsymbol{N}_{k+1}$ are computed as follows.*

**Definition III.9.** *Let: $\boldsymbol{Q_4} \triangleq \boldsymbol{Q_1} + \boldsymbol{Q_3}$, $\boldsymbol{B} \triangleq \boldsymbol{P}_{\boldsymbol{G}\boldsymbol{f}_{k+1}} \boldsymbol{\nabla}^T_{\boldsymbol{f}_k} \boldsymbol{Q_4}$, $a_S(i,j)$ be used to denote the $(i,j)^{th}$ element of any matrix $\boldsymbol{A_S}$ and the covariance matrix $\boldsymbol{P}_{k+1}$ in (16) be repartitioned as*

$$\boldsymbol{P}_{k+1} = \begin{bmatrix} \boldsymbol{P}_{\boldsymbol{f}\boldsymbol{f}_{k+1}} & \boldsymbol{P}^T_{\boldsymbol{G}\boldsymbol{f}_{k+1}} \\ \boldsymbol{P}_{\boldsymbol{G}\boldsymbol{f}_{k+1}} & \boldsymbol{P}_{\boldsymbol{G}\boldsymbol{G}_{k+1}} \end{bmatrix}, \tag{19}$$

*where $\boldsymbol{P}_{\boldsymbol{f}\boldsymbol{f}_{k+1}} \in \mathbb{R}^{5L\times 5L}$ and $\boldsymbol{P}_{\boldsymbol{G}\boldsymbol{G}_{k+1}} \in \mathbb{R}^{2\times 2}$. Then*

$$\boldsymbol{\kappa}_{k+1} = \begin{bmatrix} b(1,1) + b(2,2) \\ b(1,2) + b(2,1) \end{bmatrix},$$

*and the elements of $\boldsymbol{N}_{k+1}$ are given by:*

$$\begin{aligned} n(1,1) &= q_4(1,1)p_{GG}(1,1) + q_4(2,2)p_{GG}(2,2) \\ n(2,2) &= q_4(1,1)p_{GG}(2,2) + q_4(2,2)p_{GG}(1,1) \\ n(1,2) &= \frac{1}{2}\Big(q_4(1,1)p_{GG}(1,2) + q_4(1,1)p_{GG}(2,1) \\ &\quad + q_4(2,2)p_{GG}(1,2) + q_4(2,2)p_{GG}(2,1)\Big) \\ n(2,1) &= n(1,2). \end{aligned}$$

*Note that the time index in $\boldsymbol{N}_{k+1}$ indicates that each element $p_{GG}(\cdot,\cdot)$ corresponds to $\boldsymbol{P}_{\boldsymbol{G}\boldsymbol{G}k+1}$.*

*Proof:* By the approximate Gaussian distribution $p(\boldsymbol{y}_{k+1}|I^k)$ in Lemma III.2, and standard results from linear algebra involving matrices [49], it follows that within this scheme, (17) can be written as

$$\begin{aligned} J_{inn} = \left(\boldsymbol{h}_{k+1} - \boldsymbol{y}_{\boldsymbol{d}_{k+1}}\right)^T \boldsymbol{Q_1} \left(\boldsymbol{h}_{k+1} - \boldsymbol{y}_{\boldsymbol{d}_{k+1}}\right) + \boldsymbol{\tau}_k^T \boldsymbol{Q_2} \boldsymbol{\tau}_k \\ + \ \text{tr}\left(\boldsymbol{Q_4}\left(\boldsymbol{\nabla}_{\boldsymbol{h}_{k+1}} \boldsymbol{P}_{k+1} \boldsymbol{\nabla}^T_{\boldsymbol{h}_{k+1}} + \boldsymbol{R}_{\boldsymbol{\epsilon}}\right)\right), \end{aligned} \tag{20}$$

where $\boldsymbol{h}_{k+1}$ denotes $\boldsymbol{h}\left(\boldsymbol{x}_k, \boldsymbol{\tau}_k, \hat{\boldsymbol{z}}_{k+1}\right)$. By employing the relations in (14), (15) and (19) to expand $\boldsymbol{h}_{k+1}$, $\boldsymbol{\nabla}_{\boldsymbol{h}_{k+1}}$ and $\boldsymbol{P}_{k+1}$ respectively in (20), one is able to factorize $J_{inn}$ completely in terms of $\boldsymbol{\tau}_k$. The resulting quadratic expression is differentiated with respect to $\boldsymbol{\tau}_k$ and then equated to zero in order to determine its stationary point. This leads to (18). Moreover, the resulting Hessian matrix is given by $2\left(\tilde{\boldsymbol{G}}_k^T \boldsymbol{Q_1} \tilde{\boldsymbol{G}}_k + \boldsymbol{Q_2} + \boldsymbol{N}_{k+1}\right)$, which by the statements in Definitions III.8 and III.9 can be shown to be positive definite. This means that the dual adaptive control law specified in Theorem III.1, minimizes the selected cost function $J_{inn}$ uniquely, and the inverse term in (18) exists without exceptions. ■

**Remark III.4.** *$\boldsymbol{Q_3}$ which appears in (18) via $\boldsymbol{\kappa}_{k+1}$ acts as a weighting factor, where at one extreme, with $\boldsymbol{Q_3} = -\boldsymbol{Q_1}$, the controller completely ignores the estimates' uncertainty, resulting in HCE control, and at the other extreme, with $\boldsymbol{Q_3} = \boldsymbol{0}$, it gives maximum attention to uncertainty, which leads to cautious control. For intermediate settings of $\boldsymbol{Q_3}$, the controller strikes a compromise and operates in dual adaptive mode. It is well known that HCE control leads to large tracking errors and excessive control actions when the estimates' uncertainty is relatively high. On the other hand, cautious control is notorious for sluggish response and* control turn-off *[8], [50]. Consequently, dual control exhibits superior performance by striking a balance between the two extremes.*

**Remark III.5.** *It is interesting to note that in the HCE case,* i.e. *when $\boldsymbol{Q_3} = -\boldsymbol{Q_1}$, if one sets $\boldsymbol{Q_1} = \mathbf{I_2}$ and $\boldsymbol{Q_2} = \boldsymbol{0}$, the control law in (18) is identical to the computed-torque law in (10), with $k_d = 0$ and the dynamic functions $\boldsymbol{f}_k$ and $\boldsymbol{G}_k$ replaced by their estimates $\tilde{\boldsymbol{f}}_k$ and $\tilde{\boldsymbol{G}}_k$ respectively. This clearly confirms that the HCE approach, which characterizes the majority of adaptive controllers, treats the estimates as if they were exact, which is never the case in real-life situations as argued in Section I.*

*2) UT-based Dual Adaptive Scheme:* The EKF-based dual adaptive scheme just presented employs the EKF algorithm to address the ANN weight-tuning task. Moreover, the corresponding dual adaptive control law in (18) relies on a first-order Taylor approximation of $p(\boldsymbol{y}_{k+1}|I^k)$, as detailed in Lemma III.2. In contrast, the novel UT-based dual adaptive scheme detailed in the following paragraphs uses a specifically devised form of the UKF [42], [43] as a recursive weight-tuning algorithm, to replace the less accurate EKF algorithm of the previous scheme, and in addition employs a novel dual adaptive law that uses the UT to improve on the first-order Taylor in Lemma III.2 which leads to the EKF-based control law in (18).

As argued in [43] the UKF, originally proposed by Julier *et. al.* in [42], provides a better alternative to the well established EKF to address the problem of stochastic nonlinear estimation. Both the EKF and UKF approximate the state (or parameter) distribution by a GRV. However, while the EKF propagates the mean and covariance of this GRV through the first-order linearization of the nonlinear system, the UKF uses a minimal set of deterministically chosen sample points, termed *sigma points*, that capture completely the true mean and covariance of the GRV, and propagates them through the true nonlinear system, yielding a posterior mean and covariance that are accurate up to the second order Taylor series expansion for any nonlinearity. In contrast, the EKF is accurate only up to the first-order Taylor series expansion [43]. Moreover, the UKF is a derivative-free algorithm and as shown later in Section IV-B, it is still computationally efficient enough to be implemented on available hardware in real-time practical applications.

Starting from the MLP ANN formulation of Section III-C leading to (14), we now proceed to propose the use of an UKF algorithm in prediction mode for the real-time estimation of $\boldsymbol{z}^*_{k+1}$ as follows.

**Lemma III.3.** *In the light of (13), Definition III.6, and Assumption III.3, it follows that* $p(\boldsymbol{z}^*_{k+1}|I^k) \approx \mathcal{N}(\hat{\boldsymbol{z}}_{k+1}, \boldsymbol{P}_{k+1})$, *where* $\hat{\boldsymbol{z}}_{k+1}$ *and* $\boldsymbol{P}_{k+1}$ *are computed at each control step according to the UKF Algorithm III.2. Consequently,* $\hat{\boldsymbol{z}}_{k+1}$ *is considered to be the estimate of* $\boldsymbol{z}^*_{k+1}$ *conditioned on* $I^k$, *and* $\boldsymbol{P}_{k+1}$ *can be viewed as a measure of this estimate's uncertainty.*

*Proof:* The UKF algorithm in prediction mode, presented in Algorithm III.2, is effectively the standard UKF algorithm as stated in [43] for parameter estimation, with the difference that the measurement-update step precedes that for time-update. In addition, the time-update step is advanced by one sample to obtain $\hat{\boldsymbol{z}}_{k+1|k}$ at instant $k$. Hence, the proof of Lemma III.3 follows directly that of the UKF (additive noise version) when applied to the nonlinear stochastic state-space model in (13). ■

**Lemma III.4.** *On the basis of Lemma III.3, it follows that* $p(\boldsymbol{y}_{k+1}|I^k)$ *is approximately Gaussian with mean* $\hat{\boldsymbol{y}}_{k+1}$ *and covariance* $\boldsymbol{P}_{\boldsymbol{y}\boldsymbol{y}_{k+1}}$ *given by:*

$$\hat{\boldsymbol{y}}_{k+1} = \hat{\boldsymbol{f}}_k + \hat{\boldsymbol{G}}_k \boldsymbol{\tau}_k, \tag{22}$$

Given the previous prediction $\left(\hat{\boldsymbol{z}}_{k|k-1}, \boldsymbol{P}_{k|k-1}\right)$, denoted in short-form by $(\hat{\boldsymbol{z}}_k, \boldsymbol{P}_k)$; the following UKF algorithm (prediction mode) generates the new prediction $(\hat{\boldsymbol{z}}_{k+1}, \boldsymbol{P}_{k+1})$:

*1) Sigma-points sampling and propagation:*

$$\begin{aligned}
\mathcal{Z}_{k|k-1} &= \left[\hat{\boldsymbol{z}}_k \quad \hat{\boldsymbol{z}}_k + \left(\gamma\sqrt{\boldsymbol{P}_k}\right) \quad \hat{\boldsymbol{z}}_k - \left(\gamma\sqrt{\boldsymbol{P}_k}\right)\right] \\
\mathbb{F}_{k|k-1} &= \tilde{\boldsymbol{f}}(\boldsymbol{x}_{k-1}, \mathcal{R}_{k|k-1}),\ \mathbb{G}_{k|k-1} = \tilde{\boldsymbol{G}}(\mathcal{G}_{k|k-1}) \\
\mathbb{Y}_{k|k-1} &= \mathbb{F}_{k|k-1} + \mathbb{G}_{k|k-1}\boldsymbol{\tau}_{k-1} \\
\hat{\boldsymbol{y}}_k &= \sum_{i=0}^{2N} W_{mi}\mathbb{Y}_{i,k|k-1}
\end{aligned} \tag{21}$$

*2) Measurement update and estimate prediction:*

$$\begin{aligned}
\boldsymbol{P}_{\boldsymbol{y}\boldsymbol{y}_k} &= \sum_{i=0}^{2N} W_{ci}\left[\mathbb{Y}_{i,k|k-1} - \hat{\boldsymbol{y}}_k\right]\left[\mathbb{Y}_{i,k|k-1} - \hat{\boldsymbol{y}}_k\right]^T + \boldsymbol{R}_{\boldsymbol{\epsilon}} \\
\boldsymbol{P}_{\boldsymbol{z}\boldsymbol{y}_k} &= \sum_{i=0}^{2N} W_{ci}\left[\mathcal{Z}_{i,k|k-1} - \hat{\boldsymbol{z}}_k\right]\left[\mathbb{Y}_{i,k|k-1} - \hat{\boldsymbol{y}}_k\right]^T \\
\boldsymbol{K}_k &= \boldsymbol{P}_{\boldsymbol{z}\boldsymbol{y}_k}\boldsymbol{P}^{-1}_{\boldsymbol{y}\boldsymbol{y}_k}, \quad \boldsymbol{i}_k = \boldsymbol{y}_k - \hat{\boldsymbol{y}}_k \\
\hat{\boldsymbol{z}}_{k+1} &= \hat{\boldsymbol{z}}_k + \boldsymbol{K}_k\boldsymbol{i}_k \\
\boldsymbol{P}_{k+1} &= \boldsymbol{P}_k - \boldsymbol{K}_k\boldsymbol{P}_{\boldsymbol{y}\boldsymbol{y}_k}\boldsymbol{K}_k^T + \boldsymbol{Q}_{\boldsymbol{\rho}}
\end{aligned}$$

where: $\mathcal{Z}^T = \left[\mathcal{R}^T \quad \mathcal{G}^T\right]^T$, $\gamma = \sqrt{N+\lambda}$ , $N$ is the length of $\hat{\boldsymbol{z}}_k$, the scaling parameter $\lambda = \alpha^2(N+\kappa) - N$, constant $\alpha$ determines the spread of the sigma-points, constant $\kappa$ is a secondary scaling parameter, the UT weights are given by: $W_{m0} = \frac{\lambda}{N+\lambda}$, $W_{c0} = W_{m0} + 1 - \alpha^2 + \beta$, and $W_{mi} = W_{ci} = \frac{1}{2(N+\lambda)}$ $(i = 1, \ldots, 2N)$, and $\beta$ includes prior knowledge of the estimate's distribution.
Moreover, in the UKF framework the linear algebra operation of adding a column vector to a matrix is defined as the addition of the vector to each column of the matrix. For further details, including guidelines for selecting the UKF scaling parameters, one is referred to [43].

**Algorithm III.2:** The UKF parameter-prediction algorithm.

$$\textit{where,}\quad \hat{\boldsymbol{f}}_k = \sum_{i=0}^{2N} W_{mi}\mathbb{F}_{i,k+1|k}, \quad \hat{\boldsymbol{G}}_k = \tilde{\boldsymbol{G}}\left(\hat{\boldsymbol{g}}_{k+1}\right) \tag{23}$$

*and the covariance*

$$\begin{aligned}
&\boldsymbol{P}_{\boldsymbol{y}\boldsymbol{y}_{k+1}} = \\
&\quad \sum_{i=0}^{2N} W_{ci}\left[\boldsymbol{D}_{\boldsymbol{f}_i} + \boldsymbol{D}_{\boldsymbol{G}_i}\boldsymbol{\tau}_k\right]\left[\boldsymbol{D}_{\boldsymbol{f}_i} + \boldsymbol{D}_{\boldsymbol{G}_i}\boldsymbol{\tau}_k\right]^T + \boldsymbol{R}_{\boldsymbol{\epsilon}}
\end{aligned} \tag{24}$$

$$\textit{where,}\quad \boldsymbol{D}_{\boldsymbol{f}_i} = \mathbb{F}_{i,k+1|k} - \hat{\boldsymbol{f}}_k\ , \quad \boldsymbol{D}_{\boldsymbol{G}_i} = \mathbb{G}_{i,k+1|k} - \hat{\boldsymbol{G}}_k.$$

*Proof:* The equation of $\hat{\boldsymbol{f}}_k$ in (23) is derived by applying the UT to estimate the mean of $p\left(\tilde{\boldsymbol{f}}(\boldsymbol{x}_k, \boldsymbol{r}^*_{k+1})|I^k\right)$. The equation of $\hat{\boldsymbol{G}}_k$ in (23) is more straightforward since $\tilde{\boldsymbol{G}}_k$ is linear in the unknown parameters. Hence we simply employ the fact that $E\left\{\boldsymbol{g}^*_{k+1}\right\} = \hat{\boldsymbol{g}}_{k+1}$. To derive the equation of

$\boldsymbol{P_{yy}}_{k+1}$ in (24) one needs to advance the equation for $\boldsymbol{P_{yy}}_{k}$ in Algorithm III.2 by one sampling instant, and substitute for $\mathbb{Y}_{i,k+1|k}$ and $\hat{\boldsymbol{y}}_{k+1}$, using the relations leading to (21) in the same algorithm. ■

**Remark III.6.** *One should particularly note that in Lemma III.4, the evaluation of the approximate mean and covariance of $p(\boldsymbol{y}_{k+1}|I^k)$ are not based on a first-order Taylor approximation, as in the case of the EKF-based scheme specifically in Lemma III.2, but are generated through the more accurate method for approximating the statistics of random variables which undergo a nonlinear transformation, namely the UT [42].*

Algorithm III.2, in the light of Lemma III.3 constitutes the weight adaptation law for the novel UT-based MLP dual adaptive scheme. In addition, Lemma III.4 provides a real-time update of the probability density function $p(\boldsymbol{y}_{k+1}|I^k)$. This information is employed by the UT-based dual adaptive control law stated in the theorem below.

**Theorem III.2.** *The control law minimizing performance index $J_{inn}$ in (17), subject to the WMR dynamic model (6), Definitions III.7 and III.8, Remark III.3 and Lemmas III.3 and III.4, is given by*

$$\begin{aligned}\boldsymbol{\tau}_k &= \left(\hat{\boldsymbol{G}}_k^T\boldsymbol{Q_1}\hat{\boldsymbol{G}}_k + \boldsymbol{Q_2} + \boldsymbol{N_{GG}}_{k+1}\right)^{-1} \\ &\quad\times\left(\hat{\boldsymbol{G}}_k^T\boldsymbol{Q_1}\left(\boldsymbol{y_d}_{k+1} - \hat{\boldsymbol{f}}_k\right) - \boldsymbol{n_{Gf}}_{k+1}\right),\end{aligned} \tag{25}$$

*where*

$$\boldsymbol{N_{GG}}_{k+1} = \sum_{i=0}^{2N} W_{ci}\boldsymbol{D_G}_i^T\boldsymbol{Q_4}\boldsymbol{D_G}_i \tag{26}$$

$$\boldsymbol{n_{Gf}}_{k+1} = \sum_{i=0}^{2N} W_{ci}\boldsymbol{D_G}_i^T\boldsymbol{Q_4}\boldsymbol{D_f}_i \quad \textit{and} \quad \boldsymbol{Q_4} = \boldsymbol{Q_1} + \boldsymbol{Q_3}.$$

*Proof:* Given the approximate Gaussian distribution of $p(\boldsymbol{y}_{k+1}|I^k)$ specified in Lemma III.4, and standard results from linear algebra involving matrices [49], it follows that within this scheme, (17) can be rewritten as

$$\begin{aligned}J_{inn} &= \left(\hat{\boldsymbol{y}}_{k+1} - \boldsymbol{y_d}_{k+1}\right)^T\boldsymbol{Q_1}\left(\hat{\boldsymbol{y}}_{k+1} - \boldsymbol{y_d}_{k+1}\right) \\ &\quad + \boldsymbol{\tau}_k^T\boldsymbol{Q_2}\boldsymbol{\tau}_k + \text{tr}\left(\boldsymbol{Q_4}\boldsymbol{P_{yy}}_{k+1}\right).\end{aligned} \tag{27}$$

By substituting for $\hat{\boldsymbol{y}}_{k+1}$ and $\boldsymbol{P_{yy}}_{k+1}$ in (27), using the relations in (22) and (24) respectively, it is possible to factorize $J_{inn}$ completely in terms of $\boldsymbol{\tau}_k$. The resulting quadratic expression is differentiated with respect to $\boldsymbol{\tau}_k$ and then equated to zero in order to determine its stationary point. This leads to (25). Moreover, the resulting Hessian matrix is given by $2\left(\hat{\boldsymbol{G}}_k^T\boldsymbol{Q_1}\hat{\boldsymbol{G}}_k + \boldsymbol{Q_2} + \boldsymbol{N_{GG}}_{k+1}\right)$, which by Definition III.8 and (26) can be shown to be positive definite. This means that the UT-based dual adaptive control law specified in (25) minimizes (17) uniquely, and the inverse term in (25) exists without exceptions. ■

Remark III.4 and III.5; with (18) replaced by (25) and $\boldsymbol{\kappa}_{k+1}$ replaced by $\boldsymbol{n_{Gf}}_{k+1}$ in the former, and with $\hat{\boldsymbol{f}}_k$ and $\hat{\boldsymbol{G}}_k$ replaced by $\hat{\boldsymbol{f}}_k$ and $\hat{\boldsymbol{G}}_k$ in the latter; also apply in the context of this scheme.

## IV. Simulation and Experimental Results

As pointed out in Section I, the performance of dual adaptive controllers is typically tested by computer simulations and real-life experiments. In this section we present a number of *both* simulation and experimental results to demonstrate the effectiveness of the novel UT-based adaptive control scheme and to compare it with the EKF-based scheme originally proposed in [26] and briefly revisited in this paper.

### A. Simulation Results

Some of the parameters in our simulations namely; the measurement noise and the robot mass, inertia and friction; are programmed to change arbitrarily from one simulation trial to the other. This renders the simulations more realistic but also nondeterministic. For this reason we do not base our controller validations and comparisons on a single simulation trial, but opt to perform a Monte Carlo exercise that involves 500 simulation trials instead. To strengthen our analysis even further, we employ a statistical hypothesis test using the data acquired from the Monte Carlo simulations as detailed later in this section.

The differential WMR under study is simulated via the continuous-time dynamic model given by (1) and (5). As indicated previously, a number of parameters in this model namely $d,\ m_c,\ I_c$ and $\bar{\boldsymbol{F}}(\boldsymbol{\nu})$, are programmed to vary from one simulation trial to the other. These variations adhere to the physics of arbitrarily but realistically generated scenarios, comprising various robot load configurations and frictional conditions. Specifically, in the initialization stage of each simulation trial the modelled WMR is virtually loaded with a point mass, ranging from $0$ to $10$ kg, placed on the axis perpendicular to the driving axle at a distance, ranging from $-0.5$ to $0.5$ m, away from $\text{P}_\text{o}$. Effectively this yields a new set of arbitrary but realistic values for $d,\ m_c$ and $I_c$. Moreover, wheel viscous friction is included in the simulation by setting $\bar{\boldsymbol{F}}(\boldsymbol{\nu}) = \boldsymbol{F_c}\boldsymbol{\nu}$, where $\boldsymbol{F_c}$ is a diagonal matrix of coefficients whose values are randomly generated afresh from a uniform distribution ranging from $0.001$ to $0.5$, prior to each simulation trial. All the other WMR parameters are held constant for all simulations and are tabulated in Table I, along with the values for $d,\ m_c$ and $I_c$ that correspond to the specific case of the unloaded WMR. These parameters are based on actual measurements taken from Neurobot, the experimental WMR designed and built by the authors for the purpose of this research.

Each simulation trial consists of eight consecutive controller simulations. The first six of these correspond to the three modes of operation; *i.e.* HCE mode ($\boldsymbol{Q_3} = -\boldsymbol{Q_1}$), cautious mode ($\boldsymbol{Q_3} = \boldsymbol{0}$) and dual mode ($\boldsymbol{Q_3} = -0.8\boldsymbol{Q_1}$); for each of the two adaptive schemes being compared. The remaining two correspond to: (1) a nominally-tuned nonadaptive (NTNA) controller, which is effectively the computed-torque controller in (10) with $k_d = 0$, pre-tuned with the mean values of the

TABLE I
WMR PHYSICAL PARAMETERS (NEUROBOT WITH NO LOAD).

| Parameter | Value |
|---|---|
| $d$ | 0 m |
| $b$ | 22.95 cm |
| $r$ | 6.25 cm |
| $m_c$ | 21.0 kg |
| $m_w$ | 1.5 kg |
| $I_c$ | 0.55 kgm$^2$ |
| $I_w$ | 0.0006 kgm$^2$ |
| $I_m$ | 0.01 kgm$^2$ |

robot dynamic parameters, specifically: $\bar{d} = 0$ m, $\bar{m_c} = 26$ kg, $\bar{I_c} = 0.87$ kgm$^2$ and the diagonal values of $\bar{\boldsymbol{F_c}}$ both set to $0.25$. It is important to appreciate that this is the best a *nonadaptive* controller can do when the exact robot parameters are unknown to the controller, as in the case of these simulations and typical real-life applications; (2) a perfectly-tuned nonadaptive (PTNA) controller, which is effectively the computed-torque control law (10) with $k_d = 0$, pre-tuned with the exact values of the robot parameters. The latter is the best theoretical controller since it perfectly cancels the nonlinearities and yields *deadbeat* control. Naturally this controller is unrealistic since the exact robot parameter values are never known in practice and are generally prone to change. Hence we use this controller solely to provide an ideal reference for quantitative comparisons. In contrast, the HCE, cautious and dual adaptive controllers assume no preliminary information about the robot dynamics whatsoever, since closed-loop control is activated immediately with the initial parameter estimate vector $\hat{\boldsymbol{z}}_0$ generated randomly from a zero-mean, Gaussian distribution with variance $0.025$.

For the sake of fair comparison the same control sampling interval ($T = 50$ ms), velocity measurement noise sequence $p(\boldsymbol{\epsilon}_k) \sim \mathcal{N}(0, 0.0001\mathbf{I_2})$, reference trajectory, initial conditions, initial filter covariance matrix ($\boldsymbol{P}_0 = 0.5\mathbf{I_{27}}$), artificial process noise covariance ($\boldsymbol{Q_\rho} = 10^{-8}\mathbf{I_{27}}$), tracking error penalty ($\boldsymbol{Q_1} = \mathbf{I_2}$), and control input penalty ($\boldsymbol{Q_2} = \mathbf{0}$), are used in each controller simulation in a particular simulation trial. In addition, the sigmoidal MLP ANN used in each of the two schemes under test contains five neurons ($L = 5 \Rightarrow N = 27$). Our experiments indicated that adding more neurons did not improve the control performance significantly. In the UT-based scheme, the UKF scaling parameters are set to $\alpha = 1, \kappa = 0$ and $\beta = 2$.

*1) Single Trial Analysis:* A number of simulation results typifying the performance of the three control modes of the proposed UT-based adaptive scheme as well as the EKF-based adaptive scheme revisited in this paper are depicted in Figure 5. It should be emphasized at the outset that these results are only included to depict the typical performance of each adaptive control mode (HCE, cautious and dual) of each scheme, and not to be used to compare the two schemes (the UT-based and the EKF-based) themselves. The reason for this is that the results shown in Figure 5 correspond to single simulation trials, and since the nature of the simulation is stochastic, it is inappropriate to draw general conclusions based solely on the result of one or two simulation trials. The Monte Carlo analysis that follows later in this section is designed to address this issue and leads to a more fair and scientifically sound comparison of the proposed schemes. However, the single trial results presented in Figure 5 do give a number of important indications on the relative performance of the HCE, cautious and dual adaptive control modes, which we have found to be highly consistent and independent on the number of trials and even the scheme itself.

In Figure 5, the plots labelled (.i) correspond to the proposed UT-based scheme while those marked (.ii) correspond to the EKF-based scheme. The following comments and observations apply to both schemes. Plots (a.i) and (a.ii) depict the WMR, controlled by the respective adaptive controller in dual mode, tracking a demanding reference trajectory with nonzero initial tracking error. It is clear that the robot converges quickly to the reference trajectory and keeps tracking it with high precision, even when it reaches high speeds of around 1 m/s. Plots (b) to (e) focus on the transient of another simulation trial that uses the same reference trajectory, but purposely initiated with zero tracking error conditions. In this manner, any transient errors can be attributed to the capability of the respective controller to cope with the initially high levels of uncertainty in the estimates. Plots (b.i) and (b.ii) compare the Euclidean norm (denoted by $\|\cdot\|$ throughout the paper) of the $x - y$ position error vector. This is computed via $\|xy_{\text{error}}\| = \sqrt{(x_r - x)^2 + (y_r - y)^2}$. Plots (c.i) and (c.ii) show the magnitude of the WMR orientation error for the three control modes. Plots (d.i) and (d.ii) show the error in the robot pose while Plots (e.i) and (e.ii) compare the corresponding control inputs, more specifically the Euclidian norm of the torque vector. As can be seen in Plots (e.i) and (e.ii), the HCE controller leads to very high transient control inputs. This is a direct results of its aggressive and incautious nature, stemming from the fact that it completely ignores the high uncertainty in the initial estimates. Plots (b) to (d) clearly indicate that this leads to relatively high transient errors in both position and orientation. The cautious mode, which leads to lower transient errors relative to the HCE, is slightly more sluggish than the dual mode. This can be seen in Plots (e), where the initial control input issued by the cautious controller is the lowest. This leads to a slower (relative to the dual mode) decay of the pose error as indicated in Plots (d.i) and (d.ii). It is clear that the dual mode manages to strike a balance between these two extremes and leads to the best transient performance in both schemes. All these observations are in accordance with the anticipations of Remark III.4. In addition, the three adaptive modes in each scheme converge to the same performance at steady-state. This is not unexpected due to the fact that by the time steady-state is reached the parameter estimates would have practically converged to the actual parameters, meaning that the robot would have adapted well to its own current dynamics.

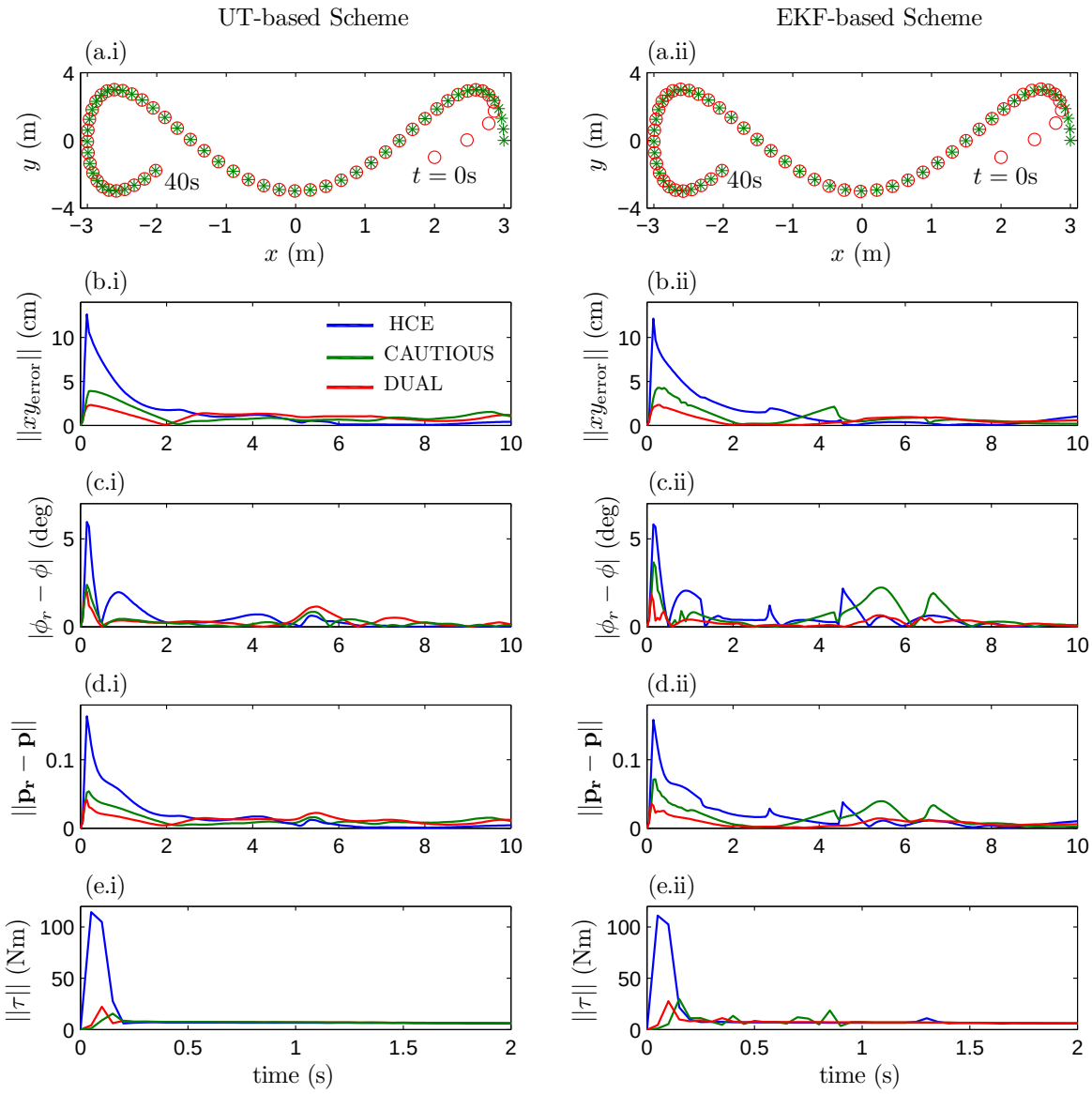



Fig. 5. Simulation results for the (i) UT-based and (ii) EKF-based schemes: (a) reference (green ×) and actual (red ◯) trajectories, (b) position error $\|xy_{\text{error}}\| = \sqrt{(x_r - x)^2 + (y_r - y)^2}$, (c) orientation error, (d) pose error, (e) control input. N.B. (a) controller in dual mode with non-zero initial error, (b) to (e) transients for zero initial error.

*2) Monte Carlo Analysis:* To quantify the controllers' performance objectively, a Monte Carlo simulation involving 500 simulation trials was performed. For each of the eight controller simulations in a trial, the reference trajectory depicted in Figure 5a, but with zero initial tracking error, is used and the simulation settings and conditions specified earlier apply. At the end of each trial, the following accumulated cost function $\mathcal{C}(k_{end})$ is calculated:

$$\mathcal{C}(k_{end}) = \sum_{k=1}^{k_{end}} \|\boldsymbol{p}_{\boldsymbol{r}k} - \boldsymbol{p}_k\|^2.$$

This cost function, based on the robot pose error over the whole time horizon ($k_{end}$ sampling instants), serves as a performance measure for each of the eight controllers operating under the same conditions, where lower values of $\mathcal{C}(k_{end})$ are naturally preferred.

The salient statistical features of the resulting eight cost distributions resulting from this Monte Carlo simulation, are depicted in the boxplot of Figure 6. Additionally, the median, interquartile range (IQR), mean and variance of each of these distributions are given in Table II. Due to the skewness of these distributions and the high number of outliers in some of the cases, the median is preferred over the mean as a measure of central tendency while the IQR is preferred over the variance as a measure of dispersion (spread). The results in Figure 6 and Table II provide the first indications how one would rank the general performance of the controllers

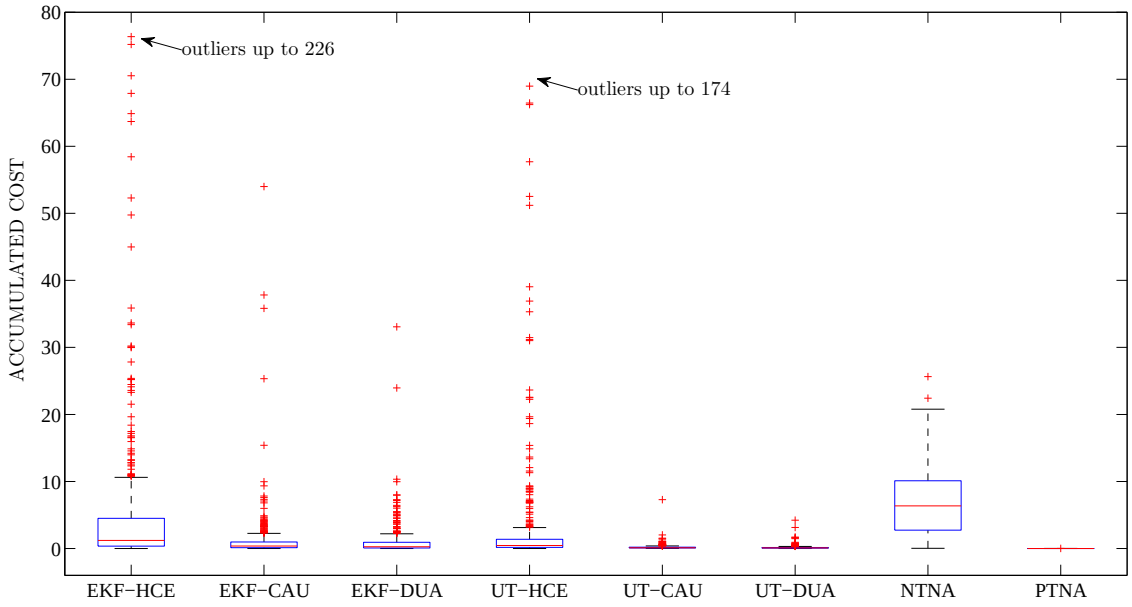



Fig. 6. Boxplot of the cost distributions.

TABLE II
STATISTICAL MEASURES OF THE COST DISTRIBUTIONS.

| Controller | Median | IQR | Mean | Variance | Rank |
|---|---|---|---|---|---|
| EKF-HCE | 1.20 | 4.16 | 7.24 | 471.12 | 6 |
| EKF-CAU | 0.37 | 0.87 | 1.18 | 14.07 | 4 |
| EKF-DUA | 0.27 | 0.85 | 0.91 | 5.06 | 3 |
| UT-HCE | 0.44 | 1.22 | 4.65 | 303.79 | 5 |
| UT-CAU | 0.12 | 0.13 | 0.19 | 0.14 | 2 |
| UT-DUA | 0.09 | 0.11 | 0.15 | 0.08 | 1 |
| NTNA | 6.36 | 7.38 | 6.86 | 21.06 | -na- |
| PTNA | 0.003 | 0.004 | 0.004 | $<$0.000 | -na- |

under investigation, where lower values of the median and IQR are obviously preferred. From the outset one can easily notice that the NTNA controller yields the highest median and IQR, implying that in general it leads to the highest pose error and deviation in performance. This is not unexpected since this controller is not adaptive and so unable to cope well with the robot parameters that are constantly changing from one simulation trial to another. In fact, its performance could be much worse if the nominal parameters, to which it is tuned, are unknown or the model variations are higher. For this reason there is no scope in comparing it further to the other adaptive controllers, and so it is withdrawn from the following comparative analysis. Consequently in the following comparative treatment we focus solely on the remaining six adaptive controllers since the PTNA results are included only for reference.

Focusing back on the six adaptive controllers, one notices that the two HCE controllers yielded a relatively high number of extreme outliers (refer to Figure 6). This is the reason why in Table II the mean and variance corresponding to these controllers are exceptionally high. This implies that in a number of trials the HCE mode led to very high transient errors. Again, this implies that the complete lack of sensitivity exhibited by HCE adaptive controllers in the face of the highly uncertain estimates characterizing the startup phase, can lead to excessively high control inputs and tracking errors which can potentially result in mission failure and possibly hardware damage in a practical situation. This strengthens our previous results in Section IV-A1 and again consolidates the arguments in Remark III.4. The results in Table II also indicate that within each scheme the dual mode outperforms the cautious and HCE modes. In addition, it is evident that each mode in the UT-based scheme outperforms its counterpart in the EKF-based scheme. The latter implies that the proposed UT-based scheme brings by a considerable improvement over the EKF-based scheme, originally proposed in [26]. However, in order to strengthen these claims further and to ascertain that the observed differences in the performance of each controller, indicated by the results in Table II, are statistically significant and cannot be attributed to chance, we employed a statistical inference procedure via the following hypothesis test.

The One-Way Analysis of Variance (ANOVA) is a powerful statistical procedure used to make inferences on the population means of several independent samples. Like all other parametric tests it relies on a number of assumptions [51]. Most importantly, the samples should be independent, normally distributed and exhibit fairly similar variances. It is also known that ANOVA is quite robust in the face of violations to its assumptions, mostly so when the sample sizes are large and equal. However, the cost distributions corresponding to the six adaptive controllers left for investigation are all positively

skewed, and therefore cannot be closely approximated to normal distributions. Hence, the original cost observations were all transformed using the natural logarithm function. The transformed samples were found to be fairly Gaussian (skewness and kurtosis in the range of $\pm 1$). This was verified by investigating the histogram and the normal quantile-quantile (Q-Q) plots of each transformed sample [51]. However, the Levene's test for homogeneity of variance [51] indicated that equal variances among the six transformed samples still could not be assumed. In such cases it is suggested that the Brown-Forsythe $F$ statistic or the Welch's $F$ statistic are used instead of the standard $F$ statistic in the ANOVA test [51].

Based on these results, the log transformed cost values were used in the ANOVA test, aimed to compare the population means of the six cost distributions. The null and alternative hypotheses for this two-tailed test are:

$H_0$ : In general the six adaptive controllers perform equally well. In other words: in an infinite number of Monte Carlo simulation trials the six controllers would yield the same mean cost.

$H_1$ : Some controllers perform better than the others. In other words: in an infinite number of Monte Carlo simulation trials two or more controllers would yield a different mean cost.

The resulting $p$-values [51], corresponding to the Brown-Forsythe and the Welch tests, were both approximately zero. Hence, since the $p$-value is smaller than the chosen level of significance $\alpha = 0.05$, the null hypothesis $H_0$ is rejected. This implies that *at least* one of the six controllers is significantly better (cost-wise) than the others. In order to investigate the underlying differences further and be able to rank the controllers according to their performance we employed the Games-Howell *post-hoc* test, which is highly recommended in the case of unequal variances [51]. The result was conspicuous, since all the $p$-values resulting from all pair-wise combinations were much lower than the chosen level of significance $\alpha$. This implies that the means of the transformed samples, depicted in Figure 7, are *all significantly different* and can be used to rank the general performance of the six adaptive controllers as given in the last column of Table II. In addition, a non-parametric test using the original cost distributions instead of the transformed distributions, namely the Kruskal-Wallis test [51] was also employed to test the set hypothesis. The final result of this analysis fully confirms that of the ANOVA.

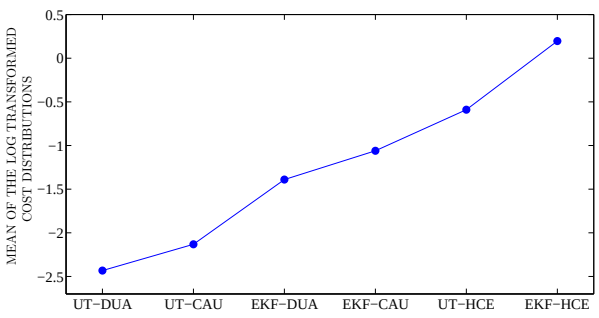



Fig. 7. Means plot of the log transformed cost distributions.

The results from this Monte Carlo comparative analysis fully support those derived from Table II and Figure 5. Hence, we can confidently claim that:

**Remark IV.1.** *The proposed UT-based scheme brings about a significant improvement in tracking performance over the EKF-based scheme, independent of the controller mode (HCE, cautious or dual). We associate this to the better estimations of the UKF over those of the EKF in the ANN training algorithm, and to the better (second-order) approximations of the UT-based control law as opposed to the first-order approximations inherent in the EKF-based control algorithm. Moreover, within each scheme the dual mode is superior to both the cautious and HCE modes. This complies with the dual control philosophy that a balance between caution and probing yields the best performance in adaptive control. It is also not surprising that the performance of the adaptive controllers is generally better than that of the computed-torque nonadaptive controller which assumes nominal values for the robot dynamic parameters, when these are prone to change.*

### B. Experimental Results

The UT-based and EKF-based dual adaptive neuro-controllers presented in this article were both implemented successfully on a physical WMR, named Neurobot, which was designed and built by the authors as an experimental research testbed. This section introduces Neurobot and reports a number of experimental results that compliment those acquired by simulation and reported in the previous section.

Neurobot, pictured in Figure 8, is a differentially driven WMR. Each of the two 125 mm diameter, solid-rubber, motorized wheels, is independently driven by a 70 W, 24 V permanent magnet dc motor (from maxon motor [52]), equipped with a 113:1 planetary reduction gearbox and a 500 pulses per revolution incremental optical encoder. Each of the two motors is driven via the LMD18200 H-Bridge IC which is controlled by a 20 kHz pulse-width modulation (PWM) reference signal. The instantaneous current in each motor is measured using the LEM HX-03-P/SP2 Hall effect current transducer, and filtered by a 4th-order continuous-time Bessel low-pass anti-aliasing filter, tuned for a corner frequency of 2 kHz, and implemented via the MAX275 filter IC. Neurobot is powered by four 12 V, 9 Ah sealed lead acid (SLA) batteries.

The algorithms controlling Neurobot were all implemented on a *MicroAutoBox* embedded computer system from *dSPACE* [53]. The *MicroAutoBox* is a compact stand-alone prototyping unit designed specifically for rapid-prototyping of computationally demanding real-time control systems, typically requiring a number of general and specialized analogue/digital input and output channels. A digital pole-placement torque controller with integral action, was designed and implemented completely in software to account for the motor electrical dynamics. This inner torque control loop uses the motor current


Fig. 8. Neurobot: the WMR built for the purpose of this research.

measurement as feedback and issues voltage commands to the motors. This ascertains that the actual torques at the wheels track those issued by the outer loop control law (the robot dynamic controller) and that motor current never exceeds a predefined safe value. This cascade approach imposes that the inner loop operates at a much faster rate than the outer loop. The sampling rates for the inner and outer loops were chosen to be 10 kHz and 200 Hz respectively.

A desktop computer was used to implement the control algorithms in *Simulink*® using the system blocks provided by the *dSpace Real-Time Interface*. *Real-Time Workshop*® is then used to automatically generate the required code which is then downloaded to the *MicroAutoBox* via the *dSpace Link Board* installed in the desktop computer. The system states and parameters along with other information about the real-time execution of each task running on the *MicroAutoBox*, such as sampling times, priorities and execution times, could also be monitored in real time via *ControlDesk*, also from *dSPACE*.

The initial network parameter vector $\hat{\boldsymbol{z}}_0$ was generated randomly. In addition, the MLP ANN contained five neurons ($L = 5 \Rightarrow N = 27$) and the UKF scaling parameters were set to $\alpha = 10^{-3}, \kappa = 3 - N$ and $\beta = 2$. The initial covariance matrix $\boldsymbol{P}_0 = 0.5\mathbf{I_{27}}$ and the process and measurement noise covariance matrices were set to $10^{-8}\mathbf{I_{27}}$ and $10^{-4}\mathbf{I_2}$ respectively. In addition, $\boldsymbol{Q_1}$ and $\boldsymbol{Q_2}$ were fixed to $\mathbf{I_2}$ and $\mathbf{0}$ respectively in all cases.

A number of experimental results, validating the proposed schemes and confirming the simulation results of this section, are presented in Figure 9. Plots (a) and (b) correspond to a challenging trajectory tracking experiment that tests the overall performance of the UT-based and EKF-based dual adaptive controllers in a real-life application. Plots (a.i) and (a.ii) show that in both cases Neurobot swiftly adapts to its own dynamics (with no preliminary offline training) and simultaneously converges smoothly to the reference trajectory, which it keeps tracking at very high precision for the rest of the experiment. Plots (b.i) and (b.ii) focus on the pose error vector norm $\|\boldsymbol{p}_{\boldsymbol{r}k} - \boldsymbol{p}_k\|$ measured during this experiment. In each case, the red trace corresponds to the dual adaptive controller while the black trace corresponds to a nonadaptive computed-torque controller subjected to the same experiment. This nonadaptive controller employs the control law in (10) with $k_d = 0$ and is tuned for Neurobot's physical parameters reported earlier in Table I. It is clear that the two dual adaptive schemes performed much better than the nonadaptive controller in steady-state. We attribute this results to the fact that the nonadaptive controller is based on a theoretical dynamic model (6), which like any other of its sort, is imperfect and relies on several physical parameters, such as friction and inertia, which are very difficult to measure precisely in practice. On the other hand, the adaptive controllers assume no preliminary information about the robot dynamics but acquire this knowledge autonomously in real-time. In addition, if one compares the pose error of the UT-based dual adaptive controller in Plot (b.i) to that of its EKF-based counterpart in Plot (b.ii), it is easy to notice that the steady-state performance of the former is relatively better than that of the latter. This result is in accordance to Remark IV.1 derived from our simulation results.

Plots (c) and (d) correspond to a different experiment with Neurobot. This experiment was designed specifically to test and compare the transient performance of the two adaptive schemes and their HCE, cautious and dual modes on a real WMR. In this experiment the reference trajectory follows a straight line along the $x$-axis, with a speed of 0.1 m/s. At $t = 5$ s, well after the robot has reached steady-state operation, the estimate vector $\hat{\boldsymbol{z}}_{k+1}$ is instantaneously reset to some randomly generated values, hence erasing all the knowledge acquired by the ANN estimator up to that point in time. In addition, the covariance matrix $\boldsymbol{P}_{k+1}$ is reset to its initial relatively high value, to reflect the high uncertainty in the new set of random network parameters. In this manner one can objectively compare the transient performance of the three control modes when faced with extremely high uncertainty in the robot dynamics. In practice, similar scenarios may arise during faults and jump variations in the robot dynamics. The question in these cases is not simply whether or not the robot adapts to the new situation, but also how smoothly and quickly it will do so. In Plots (c.i) and (c.ii), it is evident that the HCE mode (blue trace) by far yields the highest transient pose error, as a result of the sudden estimator disturbance at $t = 5$ s. As argued previously, this is clearly the result of

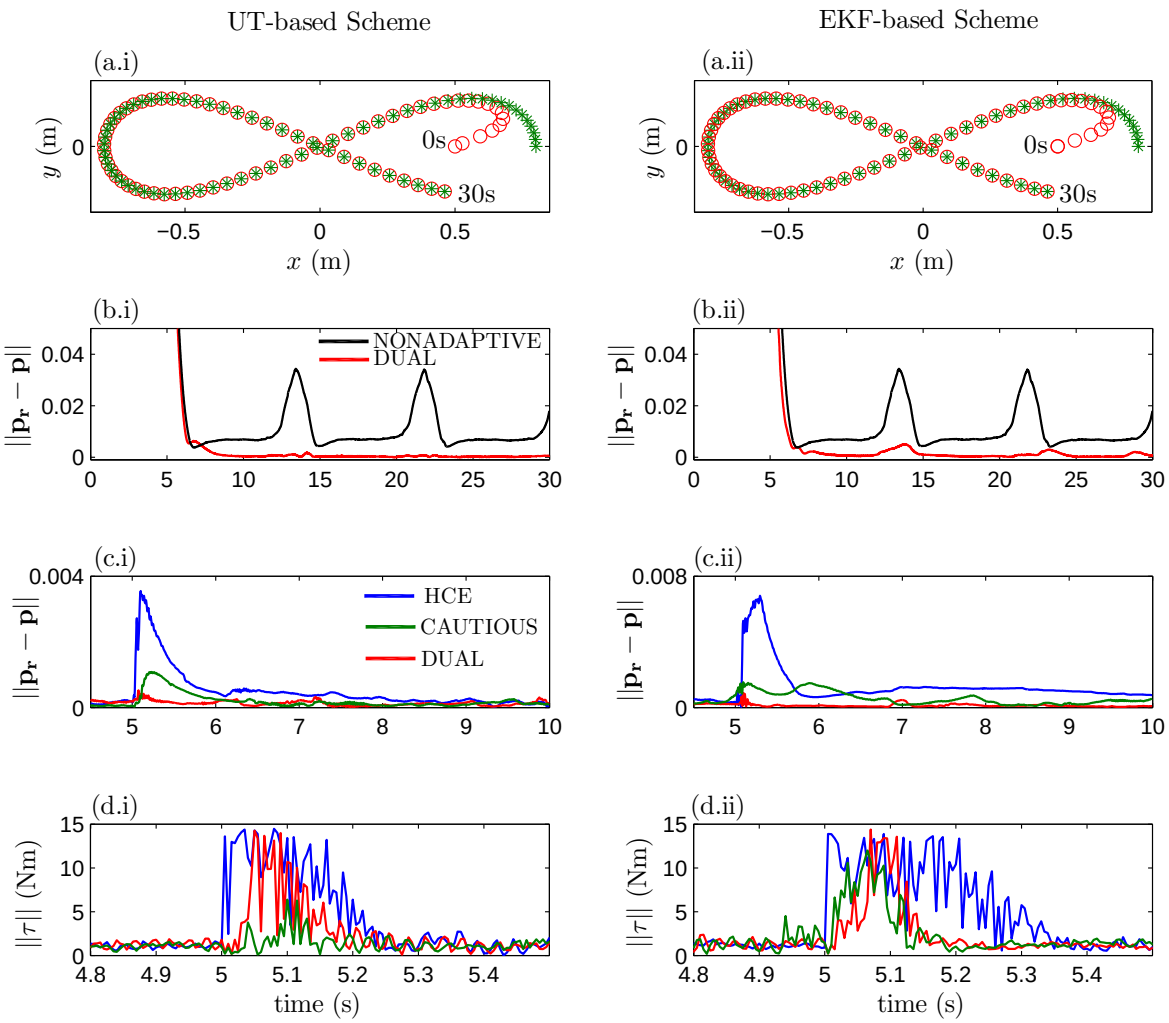



Fig. 9. Experimental results for the (i) UT-based and (ii) EKF-based schemes: (a) reference (green ×) and actual (red ◯) trajectories, (b) pose error (corresponding to the trajectory in (a)), (c) pose error (the line test), (d) control input (the line test). N.B. In (a) the controller is in dual mode (red trace) with non-zero initial error, (c) and (d) depict the line test results.

the relatively persistently aggressive and sudden control input issued by the HCE mode, which can be seen in Plots (d.i) and (d.ii) (blue trace). Specifically, these two plots depict the Euclidean norm of the *actual* torque vector developed by the motors and not that requested by the adaptive controller. In theory these are equal, but in our physical implementation we had to limit the requested torque via a saturation function so as not to damage the electronic circuitry driving the motors. If it were not for this safety feature, the situation would be closer to that depicted in Plots (e.i) and (e.ii) of Figure 5. These results also indicate that out of the three adaptive modes in each scheme, the dual mode (red traces) by far exhibits the best transient performance, due to the very low transient errors and the very quick recovery exhibited in this experiment. Moreover, it is also evident that the three controller modes in the UT-based scheme yielded lower pose errors, and hence better performance than their EKF-based counterparts. This can be clearly seen when one compares the magnitude of the pose errors depicted in Plot (c.i) with that of the errors in Plot (c.ii). One should particularly note the different scales used for the $y$-axes.

The experimental results presented in this section strongly endorse the simulation results, including those from the Monte Carlo analysis, reported previously in Section IV-A. Consequently they extend the arguments expressed in Remark IV.1 to the case of a real-life robotic application.

## V. Conclusion

In this paper we have presented a novel MLP dual adaptive control scheme for the dynamic control of WMRs. The design employs the UKF and the UT to improve on the EKF-based MLP dual adaptive scheme we recently proposed in [26]. The presented designs are validated and compared extensively via both realistic Mont-Carlo simulations, backed by rigorous

statistical analysis and real-life experiments with Neurobot, the WMR designed and built by the authors for the purpose of this research. All the results conspicuously show that:

1) The proposed UT-based scheme outperforms the EKF-based scheme.
2) In both schemes, the dual mode is superior (in transient performance) to both the cautious and the HCE controller modes.
3) The steady-state performance of the adaptive controllers is generally better than that of the computed-torque nonadaptive controller.

To the best of our knowledge this is the first time that the UT is being used in the context of dual adaptive control and where a dual adaptive controller is implemented and tested on a real mobile robot.

## REFERENCES

[1] M. K. Bugeja and S. G. Fabri, "Dual-adaptive computer control of a mobile robot based on the unscented transform," in *Proc. of the 3rd Int. Conference on Advanced Engineering Computing and Applications in Sciences (ADVCOMP' 09)*, Sliema, Malta, Oct. 2009, pp. 136–141.

[2] K. J. Åström and B. Wittenmark, *Adaptive Control*, 2nd ed. Reading, MA: Addison-Wesley, 1995.

[3] B. Wittenmark, "Adaptive dual control methods: An overview," in *5th IFAC Symp. on Adaptive Systems in Control and Signal Processing*, Budapest, Hungary, Jan. 1995, pp. 67–72.

[4] A. A. Fel'dbaum, "Dual control theory I-II," *Automation and Remote Control*, vol. 21, pp. 874–880, 1033–1039, 1960.

[5] ——, "Dual control theory III-IV," *Automation and Remote Control*, vol. 22, pp. 1–12, 109–121, 1961.

[6] ——, *Optimal Control Systems*. New York, NY: Academic Press, 1965.

[7] J. Sternby, "A simple dual control problem with an analytical solution," *IEEE Trans. Autom. Control*, vol. 21, no. 6, pp. 840–844, 1976.

[8] S. G. Fabri and V. Kadirkamanathan, *Functional Adaptive Control: An Intelligent Systems Approach*. London, UK: Springer-Verlag, 2001.

[9] N. M. Filatov and H. Unbehauen, *Adaptive Dual Control: Theory and Applications*. London, UK: Springer-Verlag, 2004.

[10] Y. Bar-Shalom and E. Tse, *Concept and Methods in Stochastic Control*, ser. Control and Dynamic Systems, C. T. Leondes, Ed. New York, NY: Academic Press, 1976.

[11] R. Milito, C. S. Padilla, R. A. Padilla, and D. Cadorin, "An innovations approach to dual control," *IEEE Trans. Autom. Control*, vol. 27, no. 1, pp. 133–137, Feb. 1982.

[12] G. A. Dumont and K. J. Åström, "Wood chip refiner control," *IEEE Control Syst. Mag.*, vol. 8, no. 2, pp. 38–43, 1988.

[13] N. M. Filatov, U. Keuchel, and H. Unbehauen, "Dual control for an unstable mechanical plant," *IEEE Control Syst. Mag.*, vol. 16, no. 4, pp. 31–37, 1996.

[14] B. J. Allison, J. E. Ciarniello, P. J.-C. Tessier, and G. A. Dumont, "Dual adaptive control of chip refiner motor load," *Automatica*, vol. 31, no. 8, pp. 1169–1184, 1995.

[15] A. Ismail, G. A. Dumont, and J. Backstrom, "Dual adaptive control of paper coating," *IEEE Trans. Contr. Syst. Technol.*, vol. 11, no. 3, pp. 289–309, May 2003.

[16] T. Tsumura, N. Fujiwara, T. Shirakawa, and M. Hashimoto, "An experimental system for automatic guidance of robot vehicle, following the route stored in memory," in *Proc. 11th Int. Symp. on Industrial Robots*, Oct. 1981, pp. 187–193.

[17] Y. Kanayama, Y. Kimura, F. Miyazaki, and T. Noguchi, "A stable tracking control method for an autonomous mobile robot," in *Proc. IEEE Int. Conference of Robotics and Automation*, Cincinnati, OH, May 1990, pp. 384–389.

[18] C. Canudas de Wit, H. Khennoul, C. Samson, and O. J. Sordalen, "Nonlinear control design for mobile robots," in *Recent Trends in Mobile Robots*, ser. Robotics and Automated Systems, Y. F. Zheng, Ed. World Scientific, 1993, ch. 5, pp. 121–156.

[19] R. Fierro and F. L. Lewis, "Control of a nonholonomic mobile robot: Backstepping kinematics into dynamics," in *Proc. IEEE 34th Conference on Decision and Control (CDC'95)*, New Orleans, LA, Dec. 1995, pp. 3805–3810.

[20] T. Fukao, H. Nakagawa, and N. Adachi, "Adaptive tracking control of a nonholonomic mobile robot," *IEEE Trans. Robot. Autom.*, vol. 16, no. 5, pp. 609–615, Oct. 2000.

[21] A. De Luca, G. Oriolo, and M. Vendittelli, "Control of wheeled mobile robots: An experimental overview," in *RAMSETE - Articulated and Mobile Robotics for Services and Technologies*, ser. Lecture Notes in Control and Information Sciences, S. Nicosia, B. Siciliano, A. Bicchi, and P. Valigi, Eds. Springer-Verlag, 2001, vol. 270, pp. 181–223.

[22] C. de Sousa, E. M. Hemerly, and R. K. H. Galvao, "Adaptive control for mobile robot using wavelet networks," *IEEE Trans. Syst., Man, Cybern.*, vol. 32, no. 4, pp. 493–504, 2002.

[23] M. L. Corradini, G. Ippoliti, and S. Longhi, "Neural networks based control of mobile robots: Development and experimental validation," *Journal of Robotic Systems*, vol. 20, no. 10, pp. 587–600, 2003.

[24] M. Oubbati, M. Schanz, and P. Levi, "Kinematic and dynamic adaptive control of a nonholonomic mobile robot using RNN," in *Proc. IEEE Symp. on Computational Intelligence in Robotics and Automation (CIRA'05)*, Helsinki, Finland, Jun. 2005.

[25] T. Das and I. N. Kar, "Design and implementation of an adaptive fuzzy logic-based controller for wheeled mobile robots," *IEEE Trans. Contr. Syst. Technol.*, vol. 14, no. 3, pp. 501–510, 2006.

[26] M. K. Bugeja and S. G. Fabri, "Dual adaptive dynamic control of mobile robots using neural networks," *IEEE Trans. Syst., Man, Cybern. B*, vol. 39, no. 1, pp. 129–141, 2009.

[27] F. Lamiraux, J. P. Laumond, C. VanGeem, D. Boutonnet, and G. Raust, "Trailer truck trajectory optimization: the transportation of components for the Airbus A380," *IEEE Robot. Autom. Mag.*, vol. 12, no. 1, pp. 14–21, 2005.

[28] R. Murphy, "Rescue robotics for homeland security," *Communications of the ACM*, vol. 27, no. 3, pp. 66–69, 2004.

[29] P. Debanne, J. V. Herve, and P. Cohen, "Global self-localization of a robot in underground mines," in *Proc. Systems, Man, and Cybernetics - Computational Cybernetics and Simulation*, Orlando, FL, Dec. 1997.

[30] M. Long, A. Gage, R. Murphy, and K. Valavanis, "Application of the distributed field robot architecture to a simulated demining task," in *Proc. IEEE Int. Conference on Robotics and Automation (ICRA'05)*, Barcelona, Spain, Apr. 2005.

[31] D. Ding and R. A. Cooper, "Electric-powered wheelchairs," *IEEE Control Syst. Mag.*, vol. 25, no. 2, pp. 22–34, 2005.

[32] R. W. Brockett, *Asymptotic Stability and Feedback Stabilisation*, ser. Differential Geometric Control Theory, R. S. Millman and H. J. Sussman, Eds. Boston, MA: Birkhäuser, 1983.

[33] I. Kolmanovsky and N. H. McClamroch, "Developments in nonholonomic control problems," *IEEE Control Syst. Mag.*, vol. 15, no. 6, pp. 20–36, 1995.

[34] P. Morin and C. Samson, "Motion control of wheeled mobile robots," in *Springer Handbook of Robotics*, B. Siciliano and O. Khatib, Eds. Berlin Heidelberg: Spinger-Verlag, 2008, ch. 34.

[35] R. Fierro and F. L. Lewis, "Control of a nonholonomic mobile robot using neural networks," *IEEE Trans. Neural Netw.*, vol. 9, no. 4, pp. 589–600, Jul. 1998.

[36] M. K. Bugeja and S. G. Fabri, "Neuro-adaptive dynamic control for trajectory tracking of mobile robots," in *Proc. 3rd Int. Conference on Informatics in Control, Automation and Robotics (ICINCO'06)*, Setúbal, Portugal, Aug. 2006, pp. 404–411.

[37] Z.-G. Hou, A.-M. Zou, L. Cheng, and M. Tan, "Adaptive control of an electrically driven nonholonomic mobile robot via backstepping and fuzzy approach," *IEEE Trans. Contr. Syst. Technol.*, vol. 17, no. 4, pp. 803–815, 2009.

[38] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. London, UK: Prentice Hall, 1999.

[39] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME J. Basic Eng.*, vol. 82, pp. 34–45, 1960.

[40] P. S. Maybeck, *Stochastic Models, Estimation and Control*, ser. Mathematics in Science and Engineering, R. Bellman, Ed. London, UK: Academic Press Inc., 1979, vol. 141-1.

[41] R. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton University Press, 1961.

[42] S. J. Julier and J. K. Uhlmann, "A new extention of the Kalman filter to nonlinear systems," in *Proc. of AeroSense: The 11th Int. Symp. on Aerospace/Defence Sensing, Simulation and Controls*, 1997.

[43] E. A. Wan and R. van der Merwe, "The unscented Kalman filter," in *Kalman Filtering and Neural Networks*, ser. Adaptive and Learning Systems for Signal Processing, Communications, and Control, S. Haykin, Ed. John Wiley & Sons, Inc., 2001, ch. 7, pp. 221–280.

[44] M. S. Radenkovic, "Convergence of the generalised dual control algorithm," *Int. J. Control*, vol. 47, no. 5, pp. 1419–1441, 1988.

[45] A. D'Amico, G. Ippoliti, and S. Longhi, "A radial basis function networks apporach for the tracking problem of mobile robots," in *Proc. IEEE/ASME Int. Conference on Advanced Intelligent Mechatronics*, Como, Italy, 2001, pp. 498–503.

[46] T.-Y. Wang and C.-C. Tsai, "Adaptive trajectory tracking control of a wheeled mobile robot via Lyapunov techniques," in *Proc. 30th Annual Conference of the IEEE Industrial Electronics Society*, Busan, Korea, Nov. 2004, pp. 389–394.

[47] Y. Yamamoto, "Control and coordination of locomotion and manipulation of a wheeled mobile manipulator," Ph.D. dissertation, Univ. of Pennsylvania, Philadelphia, USA, Aug. 1994.

[48] K. J. Åström and B. Wittenmark, *Computer Controlled Systems: Theory and Design*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 1997.

[49] K. B. Petersen and M. S. Pedersen. (2008, oct) The matrix cookbook. Version 20081110. [Online]. Available: http://www2.imm.dtu.dk/pubdb/p.php?3274

[50] N. M. Filatov and H. Unbehauen, "Survey of adaptive dual control methods," *Proc. IEE Control Theory Applications*, vol. 147, no. 1, pp. 118–128, Jan. 2000.

[51] A. Field, *Discovering Statistics using SPSS*, 2nd ed. London, UK: Sage Publications Ltd., 2005.

[52] (2010) maxon motor. [Online]. Available: http://www.maxonmotor.com

[53] (2010) dSPACE. [Online]. Available: http://www.dspaceinc.com


www.iariajournals.org

**International Journal On Advances in Intelligent Systems**
ICAS, ACHI, ICCGI, UBICOMM, ADVCOMP, CENTRIC, GEOProcessing, SEMAPRO, BIOSYSCOM, BIOINFO, BIOTECHNO, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS, CLOUD COMPUTING, COMPUTATION TOOLS
issn: 1942-2679

**International Journal On Advances in Internet Technology**
ICDS, ICIW, CTRQ, UBICOMM, ICSNC, AFIN, INTERNET, AP2PS, EMERGING
issn: 1942-2652

**International Journal On Advances in Life Sciences**
eTELEMED, eKNOW, eL&mL, BIODIV, BIOENVIRONMENT, BIOGREEN, BIOSYSCOM, BIOINFO, BIOTECHNO
issn: 1942-2660

**International Journal On Advances in Networks and Services**
ICN, ICNS, ICIW, ICWMC, SENSORCOMM, MESH, CENTRIC, MMEDIA, SERVICE COMPUTATION
issn: 1942-2644

**International Journal On Advances in Security**
ICQNM, SECURWARE, MESH, DEPEND, INTERNET, CYBERLAWS
issn: 1942-2636

**International Journal On Advances in Software**
ICSEA, ICCGI, ADVCOMP, GEOProcessing, DBKDA, INTENSIVE, VALID, SIMUL, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS, CLOUD COMPUTING, COMPUTATION TOOLS
issn: 1942-2628

**International Journal On Advances in Systems and Measurements**
ICQNM, ICONS, ICIMP, SENSORCOMM, CENICS, VALID, SIMUL
issn: 1942-261x

**International Journal On Advances in Telecommunications**
AICT, ICDT, ICWMC, ICSNC, CTRQ, SPACOMM, MMEDIA
issn: 1942-2601