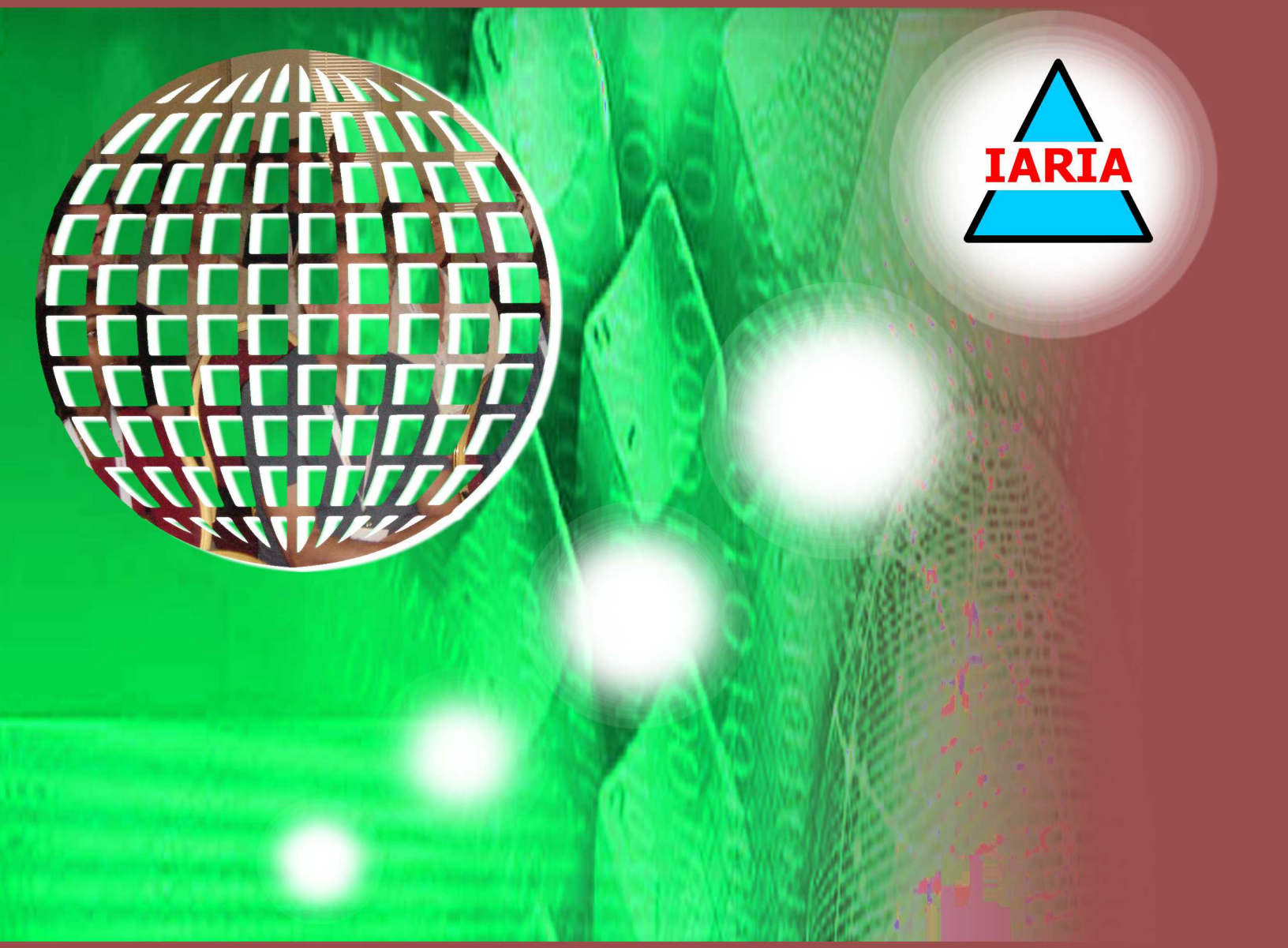
International Journal on Advances in Life Sciences



2012 vol. 4 nr. 3&4

The *International Journal on Advances in Life Sciences* is published by IARIA.
ISSN: 1942-2660
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Life Sciences
Volume 4, Number 3 & 4, 2012

International Journal on Advances in Life Sciences
Volume 4, Numbers 3 & 4, 2012

CONTENTS

Understanding Barriers to Wider Telehealth Adoption in the Home Environment of Older People: An Exploratory Study in the Irish Context

Brenda Moreira Becker Reginatto
Institute of Gerontology
King's College London
London, UK
brenda.m.b.reginatto@gmail.com

Abstract—**Emerging evidence has demonstrated the potential for Telehealth systems to reduce unnecessary hospital admission and lower costs of care by assisting patients and healthcare professionals to manage chronic conditions more efficiently. Nonetheless, due to a complex interplay of different barriers, Telehealth has not yet been widely adopted in any country. Understanding barriers to wider Telehealth adoption is vital to enable its embracement by many older people who could greatly benefit from the technology. The aim of this exploratory study was to identify barriers to wider Telehealth adoption in the homes of older people, in the Republic of Ireland. Objectives included identifying barriers from the perspective of five groups of stakeholders, determining the most pressing barriers and suggesting possible approaches to addressing such issues. Fifteen semi-structured interviews were conducted. Findings were analysed against existing literature, current technology adoption trends and successful initiatives implemented in different countries. This study suggests that the lack of incentive to healthcare professionals to embrace Telehealth, technology usability issues, implementation costs and lack of organisational willingness to change are the most pressing barriers to wider Telehealth adoption. Possible approaches to address healthcare professional incentive barriers have been suggested and include government mandates, the establishment of reimbursement schemes and the use of government financial incentives. The provision of Telehealth through devices that people are familiar with such as mobile phones, laptops and computer tablets, and the involvement of end-users during Telehealth technology development stages are also supported by this study as strategies to overcome Telehealth usability challenges.**

Keywords–older people; chronic disease management; telehealth; barriers to adoption; acceptability; incentive

I. INTRODUCTION

The original version of this paper has been presented at the eTELEMED 2012 Conference, in Valencia, Spain [1]. This extended version includes more detailed data across all sections of the paper.

In line with European demographic trends, the proportion of older people in the Republic of Ireland is expected to double in the coming decades [2], [3]. As a consequence of population ageing, Ireland is expected to experience a significant increase in the prevalence of chronic conditions. By 2020 the number of people experiencing cardiovascular disease (CVD) events is expected to rise by 50%, while the number of those diagnosed with diabetes and hypertension is likely to increase by 62% and 40%, respectively [4].

The impact of such trends on the demand and financing of healthcare services has become central in international agendas [5]. At present, chronic conditions account for three quarters of the total healthcare expenditure in Ireland. Approximately 80% of general practitioner (GP) consultations and two thirds of all emergency medical admissions to hospitals are related to chronic diseases [6]. Projections indicate that the demand for such healthcare services will continue to grow as a consequence of population ageing, representing a significant burden to the Irish public finances. The development of cost-effective and sustainable services, capable of meeting the needs of this growing population, has become essential [7].

The importance of shifting the focus of healthcare services from curative to preventative strategies, where patients are empowered to take active control over their health, is being recognised as the key to control costs and increase efficiency in healthcare [6]. It is amid this context that Telehealth technologies emerge as a relevant alternative to address these issues. Telehealth is here defined as the use of information and communication technology (ICT) based systems to assist the diagnosis, monitoring, management and empowerment of patients with chronic conditions [8]. Remote vital signs monitoring systems are a common example described in the literature [6]. In this context, Telehealth solutions allow patients to collect health measures on a regular basis, from their own homes. This information feeds into an Electronic Health Record (EHR) and is shared with a designated healthcare professional (e.g., doctor, community nurse, Telehealth triage centre) using an Internet connection. Significant changes detected in the patients' health status are brought to the attention of the healthcare provider, who may then contact the patient and intervene as necessary [9]. Telehealth features often include the use of video consultations to support remote contact between patients and healthcare professionals (ibid.). Moreover, Telehealth platforms may be also used to deliver educational content to support patient self-care [10].

Emerging evidence has demonstrated the potential for Telehealth systems to reduce unnecessary hospital admission [11], [12], decrease mortality rates [8], [13], lower costs of care per patient and increase satisfaction among users [14]. Examples of chronic conditions that may be positively supported by Telehealth include chronic heart failure – CHF [15]-[18], diabetes [19], [20] and chronic obstructive

pulmonary disease – COPD [21]-[23]. A well known example of successful Telehealth implementation is the Veterans Health Administration (VHA) Care Coordination / Home Telehealth (CCHT) scheme, in the United States. The main purpose of this programme was to coordinate the care of over 30,000 veteran patients suffering from chronic conditions (diabetes, hypertension, heart failure, COPD and depression) through the use of at-home monitoring devices, self-care tools and video consultations. Four years after its introduction, positive results have been identified including 25% reduction in number of hospital bed days used, 19% reduction in the number of hospital admissions, significantly lower costs of care per patient and a high satisfaction among users [14].

Despite all such positive factors, Telehealth has not yet been widely adopted in any country, in the sense that all relevant healthcare providers include such services within their repertoire [12], [24]. No significant trials have taken place in the Republic of Ireland and there is currently no national policy focusing on Telehealth, indicating that such systems are still far from reaching the homes of Irish older people [12]. A complex interplay of barriers has been identified in the literature and some of those have so far proven difficult to overcome [12], [24]. Poor ICT skills [12], [24], confidentiality concerns [11], [12], [25] and lack of awareness of the available technology and its potential benefits [8], [12], [24], [26], [27] were associated with lower Telehealth acceptability among both older people and healthcare professionals. Technology issues involving usability problems [8], [27], poor system stability and reliability [12], [24] have been associated with low Telehealth up-take post pilot programmes. Moreover, limited access to broadband connections [11], [12] and lack of interoperability between various Telehealth solutions have been highlighted as significant barriers to effective information sharing amongst patients and healthcare professionals [8], [10], [12], [25], [27].

The fragmentation within the healthcare sector [8], [12], [28], absence of service 'champions' capable of promoting the recognition of Telehealth as part of core healthcare services [11], [24] and overall lack of willingness to innovate [8], [11], [12], [24] have been pointed as organisational obstacles to the embracement of Telehealth in the healthcare sector. The absence of clear guidelines defining roles and responsibilities of the different stakeholders involved [11], [12], [24], [29], lack of technical quality standards [11], [24] and unclear data protection legislation are also believed to hamper Telehealth adoption amid healthcare professionals. Additionally, the lack of robust evidence supporting the role of Telehealth in chronic condition management and unclear evidence for return on investment are perceived as significant barriers to its wider adoption among the medical community [11], [12], [27].

The absence of reimbursement arrangements to incentivise healthcare providers to embrace Telehealth is perceived as a fundamental barrier to its mainstream adoption [8], [24], [26]-[28]. Additionally, it has been pointed that existent payment systems in fact discourage healthcare providers to embrace Telehealth [12], [26], [27], [30]-[32]. This is because most systems remunerate professionals per in-person contact with patients and remote contact supported by Telehealth (e.g., remote vital signs monitoring, e-mails, video-consultation) is not currently covered under most reimbursement systems.

Although much has been debated about the barriers to Telehealth adoption, little research has been done to investigate the extent to which such obstacles apply to the Irish context [12]. Moreover, few studies have attempted to explore barriers to Telehealth adoption from the perspectives of different stakeholders [29]. Therefore, the aim of the present exploratory study was to answer the following question: "what are the main barriers to the wider adoption of Telehealth in the homes of older people, in the Irish context?" Research objectives included: 1) to identify barriers to wider Telehealth adoption from the perspective of five groups of stakeholders: Potential Consumers, Healthcare Professionals, Service Providers, Technology Providers and Irish Context Experts; 2) to determine the most pressing barriers; and 3) to suggest possible approaches to address such issues.

The remainder of this paper is structured as follows: Section II explores the study methods, while Section III presents a summary of the main barriers to Telehealth adoption indentified by interviewees. Potential solutions suggested by participants are also described in this section. In Section IV findings are critically analysed against the literature, the most pressing barriers are identified and potential solutions are discussed. Study conclusion is presented in Section V, while Section VI offers a reflection upon study limitations and opportunities for further research.

II. METHODOLOGY

This study was approved by the King's College London Ethics Committee (ref KCL/10-11_379) and conducted between February and May 2011 as part of an MSc dissertation project. A maximum of fifteen semi-structured interviewees was considered feasible given the scope of the study. Potential participants were approached through convenience sampling strategy and interviewees were selected based on the assumption that they had the necessary experience to help investigating the research question.

A. Sampling and Recruitment

Based on an initial literature review, five groups of stakeholders have been defined as the subjects of this study: Consumers, Healthcare Professionals, Service Providers, Technology Providers and the Government. Since there was no provision of Telehealth services to older people in the Republic of Ireland when this study was conducted, it was not possible to verify barriers from existing consumers' point of view. For this reason, this group was replaced Potential Consumers. Additionally, despite multiple attempts, it was not possible to conduct any interviews with Irish Government representatives. In order to strengthen the analysis of barriers to wider Telehealth adoption in the Irish context, a new group was included, named Irish Context Experts. Table 1 summarizes the number of individuals invited and interviews conducted per stakeholder group.

TABLE I. NUMBER OF INDIVIDUALS INVITED AND NUMBER OF INTERVIEWS CONDUCTED PER STAKEHOLDER GROUP

Stakeholder	Number of individuals invited	Number of interviews conducted
Potential Consumers (PCs)	60*	5
Healthcare Professionals (HCPs)	4	4
Service Providers (SPs)	2	1
Technology Providers (TPs)	5	3
Irish Context Experts (ICEs)	2	2
Total		15

* Number of invitation letters made available to potential participants

To verify barriers to Telehealth adoption from the Potential Consumer (PC) point of view, relatives (sons, daughters, nephews or nieces) of older people currently receiving long-term care were approached. The rationale for selecting this group was that 1) their generations are more likely to benefit from the use of Telehealth by the time they reach old age, in comparison with their older relatives and 2) they were expected to have reasonable understanding of older peoples' needs due to their experience with relatives who require long-term care. It was assumed that this group could shed light on the research topic both from a potential user point of view and a family member / caregiver perspective. To access this group, two nursing homes in Dublin, Ireland were approached. In order to allow involvement of participants from different socio-economical backgrounds, the nursing homes chosen were located in areas that are historically known for their contrasting socio-economical differences (Nursing Home 1 was located in a less privileged area, while Nursing Home 2 situated in a more affluent region of Dublin). Invitation letters were made available at the reception desk. In order to maximise response rate, invitation was extended to visitors and staff members, who met the main inclusion criteria (sons, daughters, nephews or nieces of older people who require long-term care or suffer from chronic conditions). In total, five (n=5) PCs were recruited. Face-to-face interviews were conducted in a suitable area in the nursing homes (e.g., visitors' room).

To explore the views of Healthcare Professionals (HCPs), GPs who regularly visit residents in the same nursing homes above mentioned were approached. The reason for choosing GPs was the assumption that they 1) have relevant experience working with older patients who require chronic condition management and 2) could potentially benefit from having access to Telehealth data. An invitation letter was made available to potential participants in one of their visits to the nursing homes. In total four (n=4) HCPs were recruited. Face-to-face interviews were conducted in a suitable area in the nursing home or, alternatively, in the participant's private practice facility.

Service Providers (SPs) were defined in this study, as organisations concerned with the supply of Telecare / Telehealth products and services. Two SPs have been identified in Ireland. An invitation email introducing the study was sent to both companies. One of them (n=1) agreed to take part and a telephone interview was arranged.

Technology Providers (TPs) were defined as companies that develop Telehealth systems and have headquarters in Ireland. Five organisations have been identified and contacted through the same approach used with SPs. Three subjects (n=3) agreed to participate. Although in the case of two companies the appointed interviewee was not based in Ireland, this was considered acceptable since both individuals had the desired experience to contribute to the study. Depending on interviewees' location a face-to-face or telephone interview was arranged. Face-to-face interviews took place in a suitable area of the respondents' workplace.

Finally, Irish Context Experts (ICEs) were defined in this study as individuals who have significant knowledge of the Irish health and social care systems and are familiar with Telehealth systems. Two potential interviewees with this profile were identified through snowballing strategy (i.e., through the indication of other interviewees) and were approached via email, as described above. Both agreed to take part (n=2) and face-to-face interviews took place in a suitable area of their workplaces.

B. Data Collection and Analysis

All participants received a study information sheet and gave informed consent prior to interview. Topic guides have been used to support the semi-structured interviews and different questions have been included to suit the different stakeholders' backgrounds (available upon request from the author). A diagram created by the author illustrating possible Telehealth configurations has been used to frame discussions about barriers to Telehealth adoption (Fig. 1). The diagram displayed technologies commonly described in the literature including remote vital signs monitors, video-consultation and EHR systems. Considering the likelihood that most PCs and HCPs would not be familiar with Telehealth technologies, two videos were shown to further support the interviews. Video 1 described the use of a Telehealth remote monitoring system to support patients with chronic pulmonary disease (http://www.intel.com/corporate/healthcare/emea/eng/healthguide/LothianCaseStudy.htm). Video 2 explained the functions of a personal EHR that allows patients to organize, store, and share health information online (http://www.youtube.com/watch?v=g9hLT2bMfbY). Both videos were freely accessible on the Internet at the time the interviews were conducted and have been used for illustration purposes only. This method was considered beneficial, since interviewees expressed greater understanding of the Telehealth concept after watching the videos. Moreover, an Apple iPad device was used to display the videos. This was considered useful should interviewees be unfamiliar with touch screen interfaces, a common feature in Telehealth devices.

Interviews with SPs, TPs, and ICEs followed a similar structure. Participants were initially asked to briefly describe their background and experience in the field of Telehealth. Afterwards, the same diagram was presented to frame the

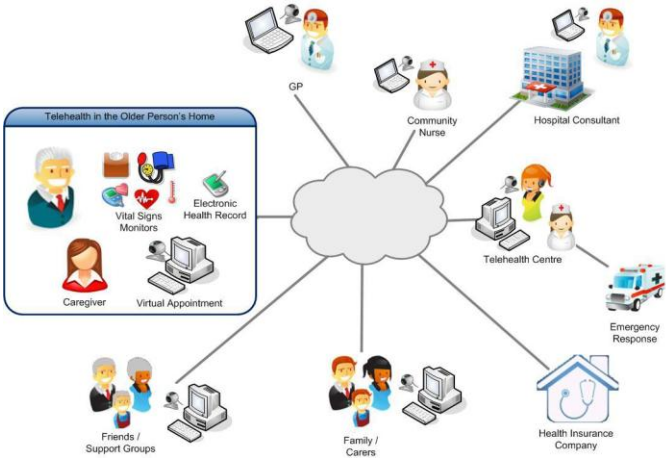


Figure 1. Diagram representing possible Telehealth configurations (larger version available upon request from the author)

discussion about the current state of Telehealth industry and existing barriers to wider implementation. Interviewees were asked to comment on whether they agreed or not with the barriers identified through the literature review. Any new barriers suggested were documented and used to frame the discussion with subsequent interviewees. At the end participants were asked to suggest solutions to overcome the barriers raised.

Interviews with PCs and HCPs were largely similar. Participants were initially asked about their familiarity with ICT solutions, as well as experience dealing with older people and chronic conditions. Figure 1 was used to illustrate possible Telehealth configurations and Videos 1 and 2 were played. Interviewees were then asked to comment on perceived advantages offered by Telehealth technologies and then potential barriers to its wider adoption in the homes of older people. Finally, participants were asked to suggest potential solutions to the barriers discussed.

Interviews were audio recorded and manually transcribed. Based on interview transcripts, a thematic content analysis was carried out. The use of the qualitative analysis software NVivo 9 (www.qsrinternational.com/products_nvivo.aspx) greatly facilitated this process.

Barriers raised by interviewees have been classified according to the categorization model described on Table 2. Categorizing the barriers for wider Telehealth adoption is important to allow for critical analysis of common aspects. Categories are also essential to identify relative criticality and therefore determine specific areas for focus and action. The classification of barriers used in this study derives from the theoretical model suggested by Broens, Veldw et al. [24]. In that model, authors classified determinants for successful Telehealth implementation under five main domains: 1) Acceptance; 2) Technology; 3) Organization; 4) Policy and Legislation and 5) Financial. In the present study, some of these categories have been subsequently broken down in order to better explore issues of greater complexity. Category 1) Acceptance in Broens, Veldw et al. model [24] has been broken down into categories A) Acceptance Barriers and E) Evidence Base Barriers. This has been done in line with other studies that identify evidence base issues not necessarily related to acceptance [12]. Moreover, one additional category has been added to the model: G) Healthcare Professional Incentive Barriers. This has been added because different studies point to a significant number of issues that have lack of incentive to healthcare professionals as its root cause [27], [31], [32], [33].

TABLE II. BARRIER CATEGORIZATION

	Broens, Veldw et al. (2007)		**Present study**
1	Acceptance	A	Acceptance Barriers
2	Technology	B	Technology Barriers
3	Organisation	C	Organisational Barriers
4	Policy and Legislation	D	Policy and Legislation Barriers
		E	Evidence Base Barriers
5	Financial	F	Financial Barriers
		G	Healthcare Professional Incentive Barriers

In order to determine the most pressing barriers to Telehealth adoption in the Irish context, findings were critically analysed and compared to the literature. Current technology adoption trends and successful initiatives implemented in different countries have been also considered in this analysis.

III. RESEARCH FINDINGS

The most significant barriers identified by interviewees are explored below. Participants also suggested possible strategies to promote Telehealth adoption, and these can be found at the end of this section.

A. *Acceptance Barriers*

1) Poor ICT skills: PCs, SP1, TPs and ICEs suggested that at present the lack of technology skills is a barrier to Telehealth adoption among older people. Nevertheless, these participants acknowledged that this should not be a barrier to Telehealth adoption in the near future. PCs believed that their generations will be more familiar with electronic devices by the time they reach old age and will have greater understanding of the advantages the technology can offer.

"My aunt had a panic button but she never wore it. I think she was afraid of the technology. I don't think older people adapt well to change. (…) I would think now we are more open, I'm only 70, I would be more open than she was. She was probably 90 when she got it. I think as the time goes on people will be more receptive to these things." (PC5)

HCPs were less optimistic than the other groups. The group pointed out that older patients often demonstrate decreased ability to learn new skills, posing an important barrier to wider Telehealth adoption. Two HCPs also believed that this will continue to be an issue for future generations of older people.

2) Lack of face-to-face contact: HCPs suggested that the lack of face-to-face contact with patients may represent an

important barrier to Telehealth adoption among medical professionals. It was pointed out that remote contact could negatively impact doctors' decision making capacity, since relying on hard data, without clinical observation, could potentially increase the risk of medical errors.

"I think baseline details like blood pressure, fine, but when you go into more details like breathlessness, wheeze, chest tightness, you can't actually see "are they cyanosed?", "what is their chest actually like?". You know, would you be able to rely on the data that much?"(HCP2)

TPs agreed that the lack of face-to-face contact may increase liability concerns among medical professionals. SP1 added that remote contact with patients may also raise fears of decreased business among physicians.

"[Doctors'] biggest drive for business is the repeated visits they receive from this demographic. So I find they are absolutely reluctant to engage with anything that may possibly reduce the amount of visits, which will happen, in their practice. That is a huge obstacle to overcome." (SP1)

On the other hand, lack of face-to-face contact has not been seen as an issue to any PCs interviewed in this study. PC1 expressed that she would rather have a video consultation with her doctor, if that would mean not having to spend time in waiting rooms before medical appointments.

3) Confidentiality concerns: PCs were interested in who would have access to their EHR. They acknowledged that different people would have different perceptions of privacy issues. While all participants would like to share their EHR with a GP and most would trust the data to a public health nurse (PHN) or competent Telehealth triage centre, there were mixed opinions about sharing Telehealth data with family members and insurance companies.

"I would be happy that my GP, my family or the nurse would be aware if I suddenly became unwell. (…) Although some people are afraid, I mean, my husband is afraid of the world and his wife knowing about his illnesses. He would feel that confidentiality might be an issue. But I think this is all to the advantage of the patient rather than the disadvantage. The advantages would certainly outweigh the disadvantages." (PC3)

Two PCs demonstrated apprehension about their health information travelling online. This was not a concern for the remainder PCs who acknowledged that, at some extent, most people already securely share sensitive information electronically (e.g., bank transactions).

Confidentiality concerns were perceived as a barrier to Telehealth implementation among HCPs. Interviewees were suspicious about how to ensure that only authorized professionals have access to EHRs and who would be ultimately liable for maintaining patient data protection.

"There are huge safety issues with having all that information accessible, and who will have access to it. Because it could just get into the wrong hands, and suddenly you are in major trouble for not protecting your patient's information. (…) At present GPs own the information to a certain extent, so if you share that with the community nurse, who owns that? And who is ultimately responsible for that if it is used inappropriately?" (HCP3)

SP1, TPs and ICEs did not believe confidentiality concerns are a significant barrier to Telehealth adoption. These groups shared the perception that this issue may be easily solved through adequate regulation and awareness raising.

4) Lack of familiarity with Telehealth and its benefits: SP1, TP1, TP2 and ICE1 agreed that the overall lack of awareness among healthcare professionals and patients about Telehealth existence and benefits is an important barrier to its wider adoption in Ireland.

"The big difficulty here in Ireland (…) in terms of the healthcare professionals is (…) the ignorance of what is the actual equipment that is out there. (…) Predominantly, they will go with what they know works, and it can be a real challenge to break that down sometimes. Telecare is much more readily accepted because the vast majority of health professionals (…) have all pretty much seen this pendant alarm so they are all pretty much familiar with that. But Telehealth is a total new department." (SP1)

B. Technology Barriers

1) Technology usability: All PCs expressed that touch screen devices, as displayed in Video 1, looked easy to use. They suggested that this may facilitate the interaction of older people with the technology, increasing the chances of adoption of Telehealth solutions. However, when asked if they would like to have a device to cater specifically for their health needs most PCs expressed that they would prefer if this could be done through the devices they already have at home, such as laptops and mobile phones. The reasons supporting these views included privacy concerns and practical issues such appliance size, mobility, ease to use and level of disruption to users' lifestyle.

"I wouldn't like to have a specific health device at home. (…) I think it is the whole thing about privacy, you know, you can have as many people on the computer and they've got their own password so it remains private." (PC2)

"I would rather have something small like that (pointing to an iPhone). Something like that would be easier to use, easier to store and you could have beside the bed, and carry around if you go out (…) I wouldn't like to have another device in the house. People don't have space for all these stuff, do they?" (PC5)

"I would like to have something in my pocket, which could do it more or less automatically. Personally I don't think people are prepared to sit down and enter information on the computer, well, I wouldn't do that. People get tired; you don't have the same drive all the time. I think if it was automatic, that would be better." (PC3)

SP1, TPs and ICEs acknowledged that existent Telehealth systems are still in early stages of development

and many issues around technology usability must be further explored. ICEs added that the frequent lack of gerontological expertise among Telehealth designers is an issue that must be addressed.

"Design challenges are huge because you are dealing with unwell people and older people. So it is much easier to get a bunch of young engineers to go crazy over the iPhone and do all kinds of this jazzy stuff (…). Somebody who is sick with COPD just needs to press the button and make it work. That is all they want to know. (…) But yet, there are innovative ways of doing that." (ICE1)

"What is missing is gerontological expertise and awareness in the technology providers. (…) They need corporate gerontologists. I think a big issue is that they are not gerontologically aware, they kind of imagine what older people are going to be." (ICE2)

Usability issues have also been identified as a barrier to Telehealth embracement by healthcare professionals. TPs suggested that the views of healthcare providers may not be sufficiently considered by technology designers, resulting in systems that are not in tune with providers' workflow.

"A lot of the technology providers in this space have already developed stuff and are trying to sell it as hard as they can (…) but what they misunderstand is that all of these things are only useful in providing data once you decided what you are trying to measure. (…) It is not about making the best possible device, it is about having the right design so the whole thing is easy to use. And it has to be clinically useful." (TP2)

HCPs also demonstrated concerns about how realistic it would be to incorporate Telehealth into their usual practice, as they may not have time to interpret the additional information generated. HCP3 and HCP4 suggested that, in order to get doctors involved, the data would have to be presented in a very simple manner, as they feel that healthcare professionals will not have the time to spend in extensive data analysis.

"I had a patient today who brought me a reading of his diabetes in a graph, so it makes it much easier to review. (…) But it can be quite time consuming, that consultation took over 30mins (…) sometimes it is just too much information, you know?"(HCP4)

"I think it would depend on the quality of the information coming through, and how user friendly it is to look at the information. (…) If you are going to be presented with pages of graphs or diagrams you are just not going to… all you want to know is 'are they getting sick or not', 'do I need to see them or not'. (HCP3)

PCs also expressed disbelief that doctors would have the time to analyse large amounts of data generated by Telehealth and suggested that many doctors may choose not to consider it when making patient-related decisions.

2) Limited access to broadband: SP1, TPs and ICEs agreed that lack of access to broadband is an important obstacle to reach older people in Ireland.

"Lack of broadband to me is probably one of the number one issues in Ireland (…). Free public broadband for older people would fix a lot of problems (…) social connectivity would increase dramatically and we could get healthcare into the homes." (TP1)

3) Device Incompatibility: TPs pointed that, at present, Telehealth devices are often not compatible with each other, what may also hamper widespread adoption. Nevertheless, it has been acknowledged that there is a significant movement pushing towards equipment compatibility, through the Continua Health Alliance (www.continuaalliance.com).

*4) Lack of interoperability amongst EHR systems:*TP3 stressed that even where the end user has the required infrastructure to run a Telehealth system at home, this is of little use if the data cannot be shared with healthcare professionals. He explained that, at present, electronic systems used by hospitals and medical practices are not technically configured to allow such data flow. Even if users were able to start collecting and recording their vital signs, unless their healthcare professional used a compatible system in their practice, they would not be able to easily share the information collected, hindering potential Telehealth capabilities. TPs and ICEs acknowledged that the industry is still very elementary, and therefore far from relying on interoperable information sharing platforms.

C. Organisational Barriers

1) Lack of integration within the healthcare sector: ICEs suggested that the actual configuration of the Irish healthcare system and lack of professionals specialised in care management could hinder Telehealth implementation in Ireland. ICE1 explained that system configurations that include care managers may allow smoother incorporation of Telehealth, thus increasing the chances of success.

"A care manager is somebody that works on behalf of the patient, and liaises with the nursing system, GP system, welfare system and admission and discharge to hospital. Most countries who have adopted Telehealth have adopted care management (…) Telehealth links different parts of the health system together. And in order to do that you need someone managing that linkage. This is one of the key issues in Ireland, we don't have that system in place." (ICE1)

2) Low levels of trust among stakeholders: PC4, SP1 and TPs suggested that low levels of trust from medical professionals in their patient's ability to measure their readings appropriately, as well as in the accuracy of devices and security of connections used may pose obstacles to wider Telehealth implementation. TP3 challenged this argument since data collected by patients should be considered as trustworthy as subjective information reported by them during medical appointments.

"What you see a lot is that the professionals can't really trust the data that is coming from the patient. I don't think that sort of barrier holds much weight. Because ultimately, when the patient walks into a doctor's office and tells them about their condition, that is no more or

less trustworthy than the patient recording it and sending. (…) I think [this attitude] is making it harder for this type of data emigration to penetrate in the industry, but I think it will go away at some point." (TP3)

As shown above, HCPs demonstrated concerns about older peoples' capacity to collect trustworthy data at home. Nevertheless, HCP3 recognized that the use of electronic devices could in fact increase data reliability, when compared to manually collected data.

"No more than the diabetic who would come back to you filled in their 'allegedly' last month glucometre readings, when you know well that they sat last night and filled them out (laughs). Whereas if it would have been done electronically, whereby there was a computer day, time and reading, which can't be altered, definitely, that would help." (HCP3)

SP1 argued that he perceives considerable reluctance from medical professionals to trust technology and service providers. He explained that Telehealth is often perceived as an invasion of the medical domain, and medical professionals are reluctant to accept that other professionals may be suggesting better ways of doing their job.

3) Increased professional responsibility and lack of organisational willingness to change: SP1, TPs and ICEs acknowledged that wider adoption of Telehealth requires healthcare professionals to significantly adapt their professional practice. Interviewees explained that, for example, Telehealth enables professionals to look after a much larger number of patients and to provide more continuous care than they would through traditional methods. It was suggested that an overall lack of willingness to embrace such changes may be a significant barrier to Telehealth implementation. TP2 explored this issue, explaining that the Telemedicine concept, for example, was more accepted because it did not fundamentally interfere in the medical practice.

"In Telemedicine (…) you are basically taking account of the fact that you have an specialist who cannot be everywhere at once, and you are using this separational distance in order to provide more use of that very scarce resource to patients, wherever they happen to be. So it fits in very well with the model of practicing medicine. As soon as you go into Telehealth it is a different matter, (…) because today's practice of medicine means that the doctor doesn't have to pay special attention to you until you are in a consultation with him. (…) If the patient is at home generating health information continuously, (…) this is a different way of practicing care, which isn't what they have been trained instinctively to do." (TP2)

Among HCPs, one interviewee clearly expressed he would not be willing to change his usual practice in order to adopt Telehealth.

"This my own perspective, I've studied medicine to deal with people, I didn't study medicine to look at their computer printouts, or blood pressure going up and down. (…) that might work, but that wouldn't be for me. I mean, it would wreck my head now if I would spend half of my day looking at printouts, or people emailing me stuff about it, I just don't do that, you know?" (HCP1)

Based on their experience dealing with medical professionals, SP1 and ICE1 argued that incongruent training is often used as a reason for not engaging with Telehealth.

"My perception is that there is almost a level of condescension, that (…) 'I am trained to be a doctor, not to look at computers, look at graphs'. And I am like, wait for a second here, this is public money that should be spent a hell of a lot smarter." (SP1)

4) Lack of 'champions' in the healthcare system: SP1, TPs and ICEs suggest that the lack of strong 'champions' in the healthcare system may be one of the reasons for the virtually inexistent movement towards Telehealth adoption in the Irish context.

D. Policy and Legislation Barriers

SP1 and ICEs argued that it is still largely unclear how data protection legislation applies to Telehealth. They suggested that this issue is an obstacle to different stakeholders to become involved with Telehealth.

Most HCPs suggested that Telehealth would not achieve wider adoption in Ireland without a clear Government led strategy. They believed that this would be necessary to address data protection concerns.

"I think [data protection legislation] would have to be determined by the government, there would have to be policies in place in terms of safety, informed consent (…) because if it is just done through the private companies I think it will be perceived as too ad-hoc or that there is something in it for the individual company." (HCP3)

E. Evidence Base Barriers

TPs and ICE1 acknowledged that despite a significant number of successful pilots, Telehealth still lacks robust studies, such as large randomized controlled trials (RCTs), to support its efficacy and cost-efficiency. Participants explained that a high proportion of Telehealth trials used small sample sizes or carefully selected participants, which greatly hinders the generalisation of results.

Nevertheless, such interviewees also pointed that the lack of RCTs may be also used as a reason for non adoption among medical professionals. It was suggested that Telehealth may instead require different scientific evaluation methods to demonstrate its value.

"(…) you can argue that [careful patient selection] is exactly what you have to do with Telehealth, that there is no point in randomly selecting people in the same way that you won't randomly give people drugs to treat their conditions. (…) I think quite often, evidence is used as an excuse for inaction rather than being the real reason why they won't invest. There are lots of other things happening in medicine that doesn't have evidence base." (TP2)

F. Financial Barriers

1) Costs of establishing the required infrastructure:

TPs and ICEs acknowledged that the high costs of establishing the necessary infrastructure, staff training, processes reconfiguration, etc., may be a major barrier to the adoption of Telehealth by healthcare systems that are already under financial strain, which is the case of the HSE, the national healthcare agency in Ireland. It was pointed, however, that this would largely depend on the level of government involvement in the implementation process.

"High costs of establishing infrastructure [is a barrier] only if the Government decides to do it. There is already lots of infrastructure out there for other reasons, we can piggyback on existing mobile networks or smart meters, and other things that are happening around us." (TP2)

HCPs also pointed out that the costs of system implementation could prevent small GP practices to engage in Telehealth. Interviewees demonstrated disbelief that wider adoption would be achieved without government financial incentive.

2) Telehealth affordability by older people: ICE1, HCPs and two PCs expressed that many older people may not have the resources to afford Telehealth if this is provided through out of pocket purchasing, therefore, hindering wider adoption.

"I just don't see it becoming a big thing if it is done privately (…) it wouldn't be standardised enough. There would be only certain people that would be able to avail of that service then." (HCP2)

"I'm sure these things are very costly, and after this recession I'm sure this is the last thing in peoples' minds, that they want to spend on things like this." (PC5)

This view was not shared by the other PCs, who would expect Telehealth services to have similar costs to Telecare or broadband services, which they judged reasonable considering the advantages offered.

All PCs would expect at least partial out of pocket financing of Telehealth. Participants expressed they would appreciate government subsidy, however they did not believe this would be realistic given the HSE financial situation.

G. Lack of Incentive to Healthcare Professional

According to SP1, TPs and ICEs the lack of clear incentives to healthcare professionals to embrace Telehealth may be one of the most significant barriers to its uptake. Interviewees suggested that, even if other obstacles are addressed, Telehealth will struggle to be widely adopted if healthcare professionals do not perceive clear advantages over traditional practice.

1) Absence of reimbursement arrangements: The fact that Telehealth is not currently covered by reimbursement arrangements was seen by HCPs, SP1, TPs and ICEs as a significant disincentive to the involvement of healthcare professionals. They argued that even though Telehealth may represent cost-savings to the wider healthcare system, healthcare professionals will be reluctant to engage unless reimbursement systems are created.

"You have to look at what incentives does a GP have to offer vital signs monitoring to his or her patients? Not much, because they are not under reimbursement systems, it doesn't exist in the HSE so it would be up to the GPs to do it privately. So they don't really have a huge incentive to do it."(ICE1)

Two HCPs also acknowledged that under the current reimbursement system in Ireland, GPs would not be compensated for monitoring Telehealth data. Both agreed that professionals would probably not engage without payment.

"I think there would have to be some sort of payment, because, say you have private patients that would be paid for the appointment, or with the medical card we would get a global fee. But the fee would have to be looked at because there is a lot more time [being spent]."(HCP4)

2) Disincentives caused by existent payment system: TPs and ICEs pointed out that, different than other technologies that have successfully penetrated in the healthcare industry, Telehealth does not fit into the existent procedure-driven model. Interviewees argued that technology diagnostic solutions, for example, clearly allowed professionals to increase their income streams, what is not the case of Telehealth. There was a common perception that unless the focus of reimbursement arrangements shifts from procedures to health outcomes, Telehealth will hardly penetrate in the healthcare system.

"The reasons why radiology was adopted so fast are quite simple. On the one hand it allowed you to generate more income, because you could get more patients through the radiology department more quickly. It also didn't interfere with the status quo and the workflow in the hospital in a fundamental way (…). So if the doctor gets paid to see you, but doesn't get paid to look after you when you are out of the room, why would they invest in it and pay attention to it? (…) I guess a lot of doctors will not like this because they prefer the system whereby you are paid by the appointment, because they can see an obvious way to increase their income, by increasing number of appointments." (TP2)

"*The answer to that is really simple. Our healthcare system is based on a model that incentivises poor health. (…) Nothing will change until we change that model." (TP3)*

H. Suggested Actions for Wider Adoption of Telehealth

PCs suggested that in order to achieve wider adoption, Telehealth technologies must be flexible enough to match different user's lifestyles and preferences. Participants added that devices should be small, portable and easy to use, what could be more easily addressed if Telehealth systems could run in devices people already own, such as mobile phones.

TPs and ICEs agreed that in the future technology developers should focus on the design of Telehealth software applications, as oppose to hardware. Overall, they highlighted that the input of gerontologists and healthcare

professionals is critical to successful Telehealth technology design.

PCs suggested different strategies to promote awareness among older people and family members. This included mass media advertisement (television, newspapers, Internet), availability of information leaflets in medical practices and the creation of a government approved website, with "neutral" and up-to-date information that could facilitate informed consumers' choice.

Most interviewees indicated that government-led Telehealth implementation could address several of the barriers discussed. Government legislation could, for example, address data protection and medical liability issues. National policies were seen as necessary to endorse a standardised adoption of Telehealth across the country and to promote educational support through undergraduate training and continuous professional development. Interviewees in all groups indicated that State provision would be important to ensure that older people with lower incomes have access to Telehealth. TPs also pointed that government initiatives could stimulate the establishment of the required infrastructure to allow Telehealth data sharing. This could include the subsidy of broadband for older people and financial incentives for the adoption of interoperable EHR systems by healthcare providers and organisations.

Considering the current Irish healthcare system configuration, ICE1 suggested that it may be easier to start the implementation of Telehealth through the secondary sector. She explained that chronic disease support teams are currently based in hospitals and Telehealth could offer cost-saving advantages for such departments.

"Actually most of the chronic disease management in Ireland is coming from the hospital, not from primary care. So that is why the hospital has a load of incentive, they are saying 'we are sending a load of these heart failure and COPD teams out to all these people, and that is expensive, it is time consuming, and they are still not able to catch people before they have an exacerbation. So if we have the infrastructure, the building, the team, why not put Telehealth into the hospital?'. So there you are, that is an incentive." (ICE1)

The establishment of reimbursement schemes has been the most suggested measure to incentivise Telehealth adoption among healthcare professionals. Considering the fact that Ireland has a public health system in place, reimbursement policies were also expected to be determined by government policies.

Finally, in order to avoid barriers associated with healthcare professional reluctance to embrace Telehealth, TPs and ICEs suggested that a possible strategy to achieve wider adoption of Telehealth in the homes of older people would be focusing on the development of Telehealth solutions that do not necessarily require healthcare professional involvement. Interviewees indicated that the platform created by Telehealth devices could be used to promote education, motivation and social support to patients and caregivers. As well as being a channel for accurate information and advice, Telehealth systems could promote treatment compliance among users through clear goal setting and feedback tools. Moreover, interviewees suggested that future developments should explore the capacity of Telehealth technologies to connect older people in equivalent disease stages and caregivers in similar situations. Participants argued that this approach could promote knowledge sharing and tackle social isolation, a frequent problem among chronic disease patients and caregivers. Additionally, this could address some of the business model and reimbursement issues previously mentioned, since Telehealth would be no longer seen as a medical device that is prescribed by a doctor, but a consumer device, which older people and family members could be interested in purchasing privately.

"I think this is the tip of the iceberg, because what we see in the market is, we have gone from 'this is a medical device and you have to comply with it every day' to 'this is your health coach, if you are dear to what it is advising you to do, your health outcomes will be better' and moving there onto 'this is actually your gateway into a support community', so it is not you and your device anymore, it is you and a community of other people, and you are sharing knowledge and experience, about what works for you. And these things instead of being isolating technologies, they actually reduce social isolation, because now you are connected. You can bring in Skype, video-conferencing with your doctor, your nurse, your family. And this is where it is getting interesting, because it is seen as no longer a thing that you have to do, but your support system. And something that you might be willing to pay for too, which is the other key factor here." (TP2)

IV. DISCUSSION

Study findings were analysed in order to determine the most pressing barriers to Telehealth adoption. Table 4 presents the barriers as identified in the introduction section of this paper and its significance assessment by each interviewee, representing a summary of interview findings. A heat map scheme has been used, where darker shades of grey indicate important issues and lighter shades point to trivial obstacles or issues that were not perceived as a barrier (Table 3). Barriers perceived as more significant are situated near the top of the table.

TABLE III. HEAT MAP COLOUR CODING

	Very important barrier
	Important barrier
	Trivial barrier
	Not a barrier
	Not mentioned

TABLE IV. BARRIERS TO WIDER TELEHEALTH ADOPTION HEAT MAP – SUMMARY OF INTERVIEW FINDINGS

Barriers	PC1	PC2	PC3	PC4	PC5	HCP1	HCP2	HCP3	HCP4	SP1	TP1	TP2	TP3	ICE1	ICE2
Lack of clear incentives to healthcare professionals															
Technology usability															
Absence of reimbursement arrangements															
Lack of organisational willingness to change															
Costs of establishing the required infrastructure															
Poor ICT skills															
Change in roles and levels of trust among stakeholders															
Confidentiality concerns															
Disincentives caused by existent payment system models															
Unclear data protection regulation															
Increased liability exposure															
Increased professional responsibility															
Lack of familiarity with Telehealth and its benefits															
Limited access to broadband															
Lack of champions in the healthcare system															
Telehealth affordability by older people															
Lack of interoperability amongst EHR systems															
Lack of established financial structure															
Fear of reduced business															
Lack of integration within the healthcare sector															
Reduced face to face contact															
Lack of standards ensuring technical quality and interoperability															
Lack of robust evidence for Telehealth effectiveness															
Inadequate evaluation methods															
Device incompatibility															
Lack of guidelines clearly defining roles and responsibilities															
Unclear cost saving evidence															
Technology stability and reliability															

In line with international literature, interviewees in this study suggested that the lack of clear incentives for healthcare professionals to engage in Telehealth is one of the most pressing barriers to its wider adoption [12], [27], [32], [33]. It has been acknowledged that the absence of reimbursement arrangements significantly discourages healthcare professionals to offer this service [8], [24], [26]-[28]. Moreover, the fact that Telehealth does not fit into the existent procedure-driven healthcare model was seen as another barrier to its penetration in the healthcare sector [12],

[26], [27], [30]-[32]. Past experiences in healthcare show that the introduction of new technologies is not an issue when its adoption model is aligned with existing incentive schemes. The rapid adoption of computed tomography and magnetic resonance imaging scanners in the healthcare sector in many countries is an example [34].

Although these barriers have so far proven more difficult to overcome, several countries have been successfully employing strategies to stimulate Telehealth adoption amongst healthcare professionals. Many countries use government mandates to achieve broad ICT adoption in the health sector. In Denmark and Norway, for example, high rates of electronic prescriptions have been achieved since the Governments made this practice mandatory for primary care providers [35].

The establishment of reimbursement structures is also considered vital to incentivise Telehealth adoption among healthcare professionals. Studies show that the reimbursement structure adopted will vary depending on the country's healthcare financing model and governments play a key role in defining this [12], [35]. In Sweden and the UK the existing (small scale) provision of Telehealth has been publicly funded. In Germany, regulatory changes have enabled Telehealth reimbursement through health insurers. In the Netherlands phone and e-mail consultations are reimbursed via fixed prices by the health insurance companies [12].

Innovative reimbursement frameworks, such as pay-for-performance schemes, are also being introduced in different countries in an attempt to shift away from procedure-driven models. In the UK around 15% of GPs' salaries is based on their performance against a set of quality measures [36].

Finally, government financial incentives have been used in Australia, Denmark, the Netherlands and the UK as effective policy tools to incentivise technology adoption among healthcare professionals [12], [35]. In the US the Medicare and Medicaid EHR Incentive Program offers financial stimulus of up to $44,000 / $63,750 to physicians and hospitals that demonstrate meaningful use of EHR systems certified by the Government. Since its implementation in January 2011, the scheme has registered over 326,000 eligible professionals and hospitals [37]. This is particularly significant considering that the fragmentation of the US American health sector is usually pointed as a barrier to the implementation of any measures in large scale [25].

Technology usability issues have been highlighted by virtually all interviewees, in agreement with several studies [8], [24], [27], [28]. It has been pointed by interviewees that devices specifically designed to cater for healthcare needs may not be well accepted by users. Leveraging devices that people already have, such as mobile phones or laptops, was pointed as a better strategy for Telehealth adoption, since people are already familiar with such devices and these fit more readily into their lifestyles. Participants added that this could make Telehealth more affordable and readily accessible to the public. This perception is in large agreement with trends towards the use of mobile platforms for Telehealth provision (e.g., smartphones, computer tablets). Projections indicate that smartphone applications will enable the mobile health industry to reach 500 million users in 2015 [38].

In line with interviewees' suggestions, other studies have also pointed that involving end-users during Telehealth technology development stages is key to solve usability problems [8], [27]. The input of gerontologists and healthcare professionals is critical to successful Telehealth technology design and will address many concerns related to life style and workflow disruption.

The lack of organisational willingness to conform to changes is considered a pressing barrier to wider Telehealth implementation [8], [11], [12], [24]. While the role of 'champions' in promoting change in the healthcare system has been acknowledged by interviewees and the literature [11], [24] achieving wider adoption of Telehealth will require more than individual leadership. May et al. [29] argue that in order to overcome intra-organisational inertia, coherent policies promoting an organisational vision are needed. According to Castro [25] strong national-level leadership has been essential to countries like Denmark, Finland, and Sweden to successfully drive and coordinate wider adoption of ICT in the health sector. The idea that comprehensive national strategies are required to address this and other pressing barriers to wider adoption of Telehealth is in line with perceptions from several interviewees.

The strategies above explored indicate that a high level of government involvement may be necessary to transform healthcare provision and allow wider adoption of Telehealth. Interviewees in this study point, however, that the high initial costs of establishing the infrastructure and incentivising healthcare professionals is an important barrier to government led Telehealth implementation in Ireland. Financial challenges have also been acknowledged by different studies [10], [24], [25], [27], [35] and partnerships between public and private sectors have been suggested as a way of overcoming such issues.

Barriers involving poor ICT literacy among older people pointed by previous studies [12], [24] were perceived by most interviewees as a trivial barrier. Recent evidence show that the interest of older people in technology has grown at a fast pace in the last decade [39], [40]. In Ireland, the percentage of people aged between 65 and 74 years accessing the Internet at least once a week has increased more than five times between 2003 and 2010 [41]. The use of the Internet for health purposes has also increased among the older population [39] indicating that this barrier may gradually become insignificant.

Other acceptance barriers mentioned in the literature such as lack of face-to-face contact [11] and confidentiality concerns [11], [12], [25] have also been cited by interviewees. However, it is possible to suggest that these issues are not significant barriers to Telehealth adoption for several reasons. According to Darkins et al. [14] patient satisfaction was significantly high among older participants in the VHA Telehealth programme and lack of face-to-face contact with healthcare professionals has not been observed as a barrier to Telehealth adoption. This may indicate that the benefits offered by the technology may outweigh such

concerns. Similarly, Castro [25] suggests that confidentiality concerns should not be a barrier to ICT penetration in the healthcare industry, considering that technical controls (e.g., encryption, electronic identification, audit logs) are available to ensure personal health data security. In Denmark, for example, health data is securely shared through an official e-health portal. In this context patients have access to this website and can easily control privacy functions, including monitoring who has accessed or modified their personal medical records. In Ireland the recently implemented National HealthLink Project (http://www.healthlink.ie/) is another example of how patient data can be securely shared over the Internet. At present this service allows hospitals and general practitioners to exchange patient information, such as laboratory results and discharge summaries. Patients, however, do not have access to this service.

Interviewees in this study challenged several authors [11], [12], [27] suggesting that lack of RCTs is not a significant barrier to Telehealth adoption. Participants argued that careful patient selection is desired to achieve Telehealth benefits, thus alternative study designs should be used to evaluate Telehealth value. This has been previously observed by other authors [24], [26]. The MAST is an example of a new model for Telehealth evaluation, which has been developed to support decision making in European countries [42]. There are indications, therefore, that evidence base barriers should not be of major significance.

Interviews supported the main findings of the literature in relation to certain technology barriers including limited access to broadband connections by older people [11], [12] and lack of integration between various Telehealth technological solutions [8], [10], [12], [25], [27]. It is important to acknowledge, however, that several initiatives are being undertaken and such barriers may not be significant in the long-term. The Irish government has implemented a national scheme, which aims to achieve extensive broadband coverage by the end of 2012 [43]. Moreover, with the increasingly fast adoption of smartphones, the native broadband Internet connection capabilities of those devices may in practice address the connectivity infrastructure requirements for Telehealth, as opposed to government-led, residence-based broadband connection programmes [44]. Similarly, interoperability issues are being tackled by both private and public sectors. Through the Continua Health Alliance, over 240 healthcare and technology companies worldwide are working together to set quality and interoperability standards for Telehealth solutions [45]. Studies by Anderson [12] and Castro [25] indicate that in the UK and Denmark government agencies are setting such standards.

V. CONCLUSION

The aim of this exploratory study was to determine the main barriers to the wider adoption of Telehealth in the homes of older people, in the Irish context. Research findings point to a number of direct and indirect obstacles, which largely correspond to those discussed in the international literature. Issues involving evidence base, technology interoperability and broadband access were not considered to be of major significance, given that important initiatives are already addressing these barriers. Similarly, based on interviewees' perceptions and technology adoption trends, older people's acceptance is not believed to be a pressing barrier to wider Telehealth implementation in the medium term.

The findings indicate that technology usability issues may significantly hinder Telehealth adoption. The use of devices that people are familiar with such as mobile phones, laptops and computer tablets, and the involvement of end-users during Telehealth technology development stages are supported by this study as strategies to overcome Telehealth usability challenges. Another important barrier is the lack of organisational willingness to change, currently perceived in the healthcare sector. While implementation costs were seen as a challenge to government action in Ireland, strong national-level leadership is considered essential.

Finally, the lack of incentive to healthcare professionals to embrace Telehealth is considered a pressing barrier to its wider adoption. The absence of arrangements to reimburse healthcare providers and the incongruence with the present procedure-driven healthcare model are believed to significantly discourage professionals to offer this service. Possible approaches to address healthcare professional incentive barriers have been suggested and include government mandates, the establishment of reimbursement schemes and the use of government financial incentives.

The imminent population ageing and epidemiologic trends indicate that new forms of healthcare provision are urgently needed. Shifting away from the current disease-centric healthcare model towards a health-centric system is not only an economic necessity but also a moral obligation. The adoption of Telehealth technologies is believed to contribute towards these issues by allowing more efficient service provision in a patient-driven model. However, the disruption to traditional healthcare practices caused by the introduction of Telehealth represents a major challenge, one that requires the support from all stakeholders involved. Governments around the world are gradually implementing strategies to promote a new vision in the healthcare sector and significant changes are envisaged in the long-term.

Meanwhile, technology adoption and the use of the Internet for health purposes are growing among all age groups, including the older population. Concomitantly, projections suggest a significant growth in the mobile health industry in coming years, which may enable Telehealth diffusion through a consumer market. Should these trends continue and become reflected in people's expectations around healthcare provision it is possible that, in the short-term, technological disruption in the healthcare sector will be a demand raised by patients themselves.

VI. STUDY LIMITATIONS AND AREAS FOR FUTURE RESEARCH

Even though interview findings indicate large agreement with the literature, the small scope of this study and its non-probability sampling strategy do not allow for the generalisation of these results. It is possible to suggest, however, that the comprehensive literature review and

adequate methodology employed may provide a useful framework for future research projects.

This study would have greatly benefited from the participation of representatives of the Irish Government, as initially intended. It is also important to acknowledge the possibility of biased results, since the views of those participants who volunteered to take part in this study may differ from others [46]. Additionally, the fact that Potential Consumers and Healthcare Professionals interviewed in this study did not have personal experience with Telehealth technologies indicate that one has to cautiously consider their contributions.

While identifying barriers to wider Telehealth adoption is an important starting point to promote its implementation, future research should focus on clear and practical strategies to increase Telehealth adoption in the homes of older people. Investigating innovative ways of incentivising healthcare professionals and organisations to embrace Telehealth in their usual practices seems to be one of the most needed areas for research.

ACKNOWLEDGMENT

The author would like to thank Professor Anthea Tinker from the Institute of Gerontology, King's College London, for her support and supervision.

REFERENCES

[1] B. Reginatto, "Addressing Barriers to Wider Adoption of Telehealth in the Homes of Older People: An Exploratory Study in the Irish Context," Proc. eTELEMED 2012: The Fourth International Conference on eHealth, Telemedicine, and Social Medicine. IARIA, Jan. 2012, pp. 175-183, ISBN: 978-1-61208-179-3.

[2] M.C.K.Gaßner and M. Conrad, "ICT enabled independent living for elderly – A status-quo analysis on products and the research landscape in the field of Ambient Assisted Living (AAL) in EU-27," Berlin: Institute for Innovation and Technology as part of VDI/VDE Innovation + Technik GmbH, 2010.

[3] P. McGill, "Illustrating Ageing in Ireland North and South: Key Facts and Figures," Belfast: Centre for Ageing Research and Development in Ireland, 2010.

[4] K.P. Balanda, S. Barron, L. Fahy, and A. Mclaughlin, "Making chronic conditions count: hypertension, stroke, coronary heart disease, diabetes. A systematic approach to estimating and forecasting population prevalence on the island of Ireland," Dublin: Institute of Public Health in Ireland, 2010.

[5] WHO, "2008-2013 Action Plan for the Global Strategy for the Prevention and Control of Noncommunicable Diseases," World Health Organization: Geneva, 2008.

[6] DoHC, "Tackling chronic disease - a policy framework for the management of chronic diseases," Dublin: Department of Health and Children, 2008.

[7] R. Layte, "Projecting the Impact of Demographic Change on the Demand for and Delivery of Healthcare in Ireland," Dublin: ESRI, 2009.

[8] COCIR, "COCIR Telemedecine toolkit. For a better deployment and use of Telehealth," Brussels: European Coordination Committee of the Radiological, Electromedical and Healthcare IT Industry, 2010.

[9] Counsel and Care, "Telecare and telehealth: what it is and how to get it," Counsel and Care: London, 2010.

[10] S. Koch, "Home telehealth--current state and future trends," International journal of medical informatics, 2006. 75(8): pp. 565-576.

[11] N. Goodwin, "The State of Telehealth and Telecare in the UK: Prospects for Integrated Care," Journal of Integrated Care, 2010. 18(6): pp. 3-10.

[12] L. Kubitschke and K. Cullen, "ICT & ageing–European study on users, market and technologies. Final report," Brussels: Commission of the European Communities, 2010.

[13] J. Barlow, D. Singh, S. Bayer, and R. Curry, "A systematic review of the benefits of home telecare for frail elderly people and those with long-term conditions," Journal of telemedicine and telecare, 2007. 13(4): pp. 172.

[14] A. Darkins, P. Ryan, R. Kobb, L. Foster, E. Edmonson, B. Wakefield, and A. E. Lancaster, "Care coordination/home telehealth: the systematic implementation of health informatics, home telehealth, and disease management to support the care of Veteran patients with chronic conditions," Telemedicine and e-Health, 2008. 14(10): pp. 1118-1126.

[15] R. A. Clark, S. C. Inglis, F. A. Mcalister, J. G. F. Cleland, and S. Stewart, "Telemonitoring or structured telephone support programmes for patients with chronic heart failure: systematic review and meta-analysis,". Bmj, 2007. 334 (7600): pp. 942.

[16] J. G. F. Cleland, A. A. Louis, A. S. Rigby, U. Janssens, and A. H. M. M. Balk, "Noninvasive Home Telemonitoring for Patients With Heart Failure at High Risk of Recurrent Admission and Death: The Trans-European Network-Home-Care Management System (TEN-HMS) study," Journal of the American College of Cardiology, 2005. 45(10): pp. 1654-1664.

[17] O. Dar, J. Riley, C. Chapman, S. W. Dubrey, S. Morris, S. D. Rosen, M. Roughton, and M. R. Cowie, "A randomized trial of home telemonitoring in a typical elderly heart failure population in North West London: results of the Home-HF study," European journal of heart failure, 2009. 11(3): pp. 319.

[18] B. Maric, A. Kaan, A. Ignaszewski, and S. A. Lear, "A systematic review of telemonitoring technologies in heart failure," European journal of heart failure, 2009. 11(5): pp. 506.

[19] T. E. Barnett, N. R. Chumbler, W. B. Vogel, R. J. Beyth, H. Qin, and R. Kobb, "The effectiveness of a care coordination home telehealth program for veterans with diabetes mellitus: a 2-year follow-up," American Journal of Managed Care, 2006. 12(8): pp. 467.

[20] M. Jaana and G. Paré, "Home telemonitoring of patients with diabetes: a systematic assessment of observed effects," Journal of evaluation in clinical practice, 2007. 13(2): pp. 242-253.

[21] M. Vitacca, L. Bianchi, A. Guerra, C. Fracchia, A. Spanevello, B. Balbi, and S. Scalvini, "Tele-assistance in chronic respiratory failure patients: a randomised clinical trial," European Respiratory Journal, 2009. 33(2): pp. 411.

[22] J. C. A. Trappenburg, A. Niesink, G. H. De Weert-Van Oene, H. Van Der Zeijden, R. Van Snippenburg, A. Peters, J. W. J. Lammers, and A. J. P. Schrijvers, "Effects of telemonitoring in patients with chronic obstructive pulmonary disease," Telemedicine and e-Health, 2008. 14(2): pp. 138-146.

[23] P. Koff, R. H. Jones, J. M. Cashman, N. F. Voelkel, and R. Vandivier, "Proactive integrated care improves quality of life in patients with COPD," European Respiratory Journal, 2009. 33(5): pp. 1031.

[24] T.H.F. Broens, R. M. H. A. H. I. T. Veldw, M. M. R. Vollenbroek-Huttenw, A. T. V. Hermenswz, H. J., Halteren, and L. J. M. Nieuwenhuis, "Determinants of successful telemedicine implementations: a literature study," Journal of telemedicine and telecare, 2007. 13(6): pp. 303.

[25] J.G. Anderson, "Social, ethical and legal barriers to e-health," International journal of medical informatics, 2007. 76(5-6): pp. 480-483.

[26] A. Runge and F. Feliciani, "Telemedicine: from technology demostrations to sustainable services," in The European Files: The Telemedicine challenge in Europe. Brussels: European Commission, 2010.

[27] M. Alwan and J. Nobel, "State of Technology in Aging Services According to Field Experts and Thought Leaders," Washington: American Association of Homes and Services for the Aging (AAHSA), 2008.

[28] J.-M. Pique, "Impact on the restructuring of healthcare," in The European Files: The Telemedicine challenge in Europe. Brussels: European Commission, 2010.

[29] C. May, T. Finch, J. Cornford, C. Exley, C. Gately, S. Kirk, et al., "Integrating telecare for chronic disease management in the community: What needs to be done?" BMC Health Services Research, 2011. 11(1): pp. 131.

[30] J. Bonfini and C. Parker, "Telehealth: Making House Calls Count," in The European Files: The Telemedicine Challenge in Europe. Brussels: European Comission, 2010.

[31] J. Cruickshank, "Healthcare without Walls: A Framework for Delivering Telehealth at Scale," London: 2020health.org, 2010.

[32] L.L. Barrett, "Healthy@ Home," Washington: AARP Foundation, 2008.

[33] D.W. Bates, "Physicians and ambulatory electronic health records," Health Affairs, 2005. 24(5): pp. 1180.

[34] E.H. Oh, Y. Imanaka, and E. Evans, "Determinants of the diffusion of computed tomography and magnetic resonance imaging," International journal of technology assessment in health care, 2005. 21(1): pp. 73-80.

[35] D. Castro, "Explaining International IT Application Leadership: Health IT," Washington: The Information Technology and Innovation Foundation, 2009.

[36] NHS, "GP Payments Calculation Service Memorandum of Information (MOI)," 2010 [Online]. Available from: http://www.connectingforhealth.nhs.uk/systemsandservices/gpsupport/gppcs/gppcsmoi.pdf [Accessed 09 December 2012].

[37] CMS, "EHR Incentive Programs" 2012 [Online]. Available from: http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/index.html [Accessed 09 December 2012].

[38] E. Mikalajunaite, "500m people will be using healthcare mobile applications in 2015," 2010 [Online]. Available from: http://www.research2guidance.com/500m-people-will-be-using-healthcare-mobile-applications-in-2015/ [Accessed 09 December 2012].

[39] K. Cullen, C. Dolphin, S. Delaney, and M. Fitzpatrick, "Survey of Older People and ICTs in Ireland (2008)," Dublin: Work Research Centre (WRC) & Age Action Ireland, 2009.

[40] H. Seybert and A. Loof "Internet usage in 2010 – Households and Individuals," Eurostat, European Commission, 2010.

[41] Eurostat, "Individuals regularly using the Internet," 2011 [Online]. Available from: http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=isoc_pibi_use&lang=en [Accessed 09 December 2012].

[42] K. Kidholm, C.D. Pedersen, J. Rasmussen, L.K. Jensen, A.G. Ekeland, A. Bowes, et al., "A new model for assessment of telemedicine—MAST," International Journal of Integrated Care, 2011. 11(6).

[43] DCENR, "Rural Broadband Scheme Announced by Minister Rabbitte," 2012 [Online]. Available from: www.dcenr.gov.ie/NR/rdonlyres/732A18F8-2247-4F96-8334-F63171B1F36D/0/PressReleaseRBS952011.doc [Accessed 09 December 2012].

[44] Miscrosoft Tag, "The Growth of Mobile Marketing and Tagging," 2011. [Online]. Available from: http://tag.microsoft.com/community/blog/t/The_Growth_of_Mobile_Marketing_and_Tagging.aspx [Accessed 09 December 2012].

[45] Continua Health Alliance, "About the Alliance," 2010 [Online]. Available from: http://www.continuaalliance.org/about-the-alliance.html [Accessed 09 December 2012].

[46] A. Bryman, Social Research Methods. Oxford: Oxford University Press, 2008.

Trust in Ambient Assisted Living (AAL) - A Systematic Review of Trust in Automation and Assistance Systems

Frederick Steinke
Humboldt University of Berlin
Berlin, Germany
steinkef@student.hu-berlin.de

Tobias Fritsch
Heidelberg University
Heidelberg, Germany
T.Fritsch@gmx.net

Lina Silbermann
Nürtingen-Geislingen University
Geislingen, Germany
lina.silbermann@web.de

***Abstract*—The aim of the study was the investigation of the existing literature dealing with trust in Ambient Assisted Living (AAL). Additionally, a definition of trust in AAL was derived. For that purpose, a numeric analysis of articles considering the factor trust in automation, as well as assistive technologies for older people was carried out. A systematic literature review with a total of 150 dissimilar keyword-combinations, based on three different descriptors in three bibliographic online databases, was performed. This review revealed that 18 articles deal with trust in healthcare or assistance systems, but several of them only superficially. Despite the increasing market relevance in the last decade, none of the identified studies focused explicitly on trust in AAL. As can be seen from the results, older people as a target group for qualitative and quantitative research in this field are detected, but only partially examined. For obtaining access to older persons' trust in automation in general and AAL in particular, further research is needed. To identify influencing factors on trust in AAL, a broader survey and experiments with persons over the age of 60 years should be conducted.**

Keywords-AAL; Ambient Assisted Living; Assistance Systems; Elderly People; Trust

I. INTRODUCTION

As a result of demographic change, the number of people in advanced age, who want to spend a self-determined, independent life at home, is growing. Unfortunately, not all elderly people are able to reach this goal without assistance. This often leads to conflicting goals. An age-related decline in physical fitness, as well as physical limitations in consequence of diseases or accidents, mean that the elderly need support in realizing their desire to live in their familiar surroundings. This results in tension between the affordability of traditional personal care and specific individual support, as well as novel technical support. As seen in [1] the concept of trust in Ambient Assisted Living (hereinafter AAL) should be ascribed greater importance.

On one hand, human assistance in activities of daily living (ADL) like taking a bath, preparing meal or going for a walk is a great relief for people with health restrictions. On the other hand, science has, for several decades, dealt with research into new technologies to support people in their own home [2]. Meanwhile, innovations in the home environment offer numerous opportunities for technology-supported systems. Researchers have developed a plurality of services combined with technical support for elderly people. Terms as 'Smart House' [3], 'Smart Home' [4], 'Assistive Technology (AT)' [5] or 'Ambient Assisted Living (AAL)' [6] are just a few of the frequently used terms in this context.

In the present article the importance of trust in AAL for elderly people is the focal point of interest. In order to take advantage of AAL technology, which assists an impaired person in everyday life [7], the user must have trust in this assistance system. Since, in case of an emergency, this assistance can save lives, it is obvious that the concept of trust has fundamental importance to the consideration of development, purchase and use of AAL. The fact that older people have typically grown up without technologies like personal computers, smart phones or the Internet, which are often integrated in AAL [8], implies special demands towards the design of these devices.

The present study is structured as follows: Firstly, the background section explains the development of AAL as a result of demographic change. Next, the importance of trust as an influencing factor in this research context will be highlighted and a definition for trust in AAL derived. In the third section the literature review as research framework is described in detail. The acquired data is then analyzed in its entirety and moreover, studies regarding trust in healthcare and assistance systems are considered separately. Finally, a discussion of the observed results and an overview about further research activities is provided. The design of the present study is oriented to answer the research question: *Why is the investigation of trust in AAL necessary, and what is the current state of research into it?*

II. BACKGROUND

This section contains the background information about the development of AAL as a reaction to demographic change, as well as the importance of the concept of trust as an influencing factor for AAL. In conclusion, the development of a definition of trust in AAL is detailed.

A. Development of AAL as Reaction to the Demographic Change

According to the United Nations Department of Economic and Social Affairs (UNDESA), as compared to the total population, the proportion of people over the age of 60 years is constantly increasing [9]. The number of people over the age of 60 is predicted to grow from more than 700 million in 2009, to 2 billion in the year 2050. Worldwide this would correspond to a tripling of the cohort in a period of 40 years. The annual growth rate of 'generation 60plus' amounts 2.6 percent. This increase eclipses the overall population's growth rate of 1.2 percent per annum. At the present time, over a fifth of the population in the more developed regions is 60 years of age or over. Projections indicate that nearly one third of the total population will belong to that age group by 2050 [9]. Based on this development, health care expenditure, for example within the European Union and Norway, will change dramatically [10].

Additionally, it should be mentioned that age is not readily defined in reference only to the date of birth. Although the chronological age of two persons could be equal, the biological, psychological or social age may differ [11]. Also the cohort effect may influence differences in persons' age [12]. As defined by the World Health Organization "there is no United Nations standard numerical criterion, but the UN agreed cut-off is 60+ years to refer to the older population [11]." The terminology 'older person' or 'elderly person' is used interchangeably; therefore, these terms are similarly used in the current study for people over the age of 60. The above presented facts, in combination with the existing older persons' purchasing power, accentuate the enormous importance of the elderly for science and economy.

Moreover, technological progress and a high degree of information technology are factors that are gaining more and more relevance in everyday life. The beginning of research into the field of Assistive Technology (AT) can be traced back to the early 1970's. Then, so called "phone-chains" used the standard telephone system and were organized by a network of elderly-persons and professionals [2]. Mutual telephone calls were used to monitor the group, and if a member did not respond, their doctor or relatives were notified. This can be regarded as the first electronic emergency system for elderly persons.

The next step was the development of home emergency call systems. One of the most famous was the HTS831, which had two different buttons: one red, and one green. This system consisted of a wireless transmitter, which the user was able to wear around their neck. In case of emergency the user could either push the button on the transmitter or the red button at the station, to contact the emergency center. As a security and monitoring function, the user had to press the green button once a day [2]. In the middle of the 1990's, the first video conference system for private homes was offered. TV-top boxes, or a separate video telephone, functioned as the user interface. Additionally, this system contained functions for personal discussions and organization of, for example, nursing, medical or entertainment services [13]. In summary, efforts to develop useful and coherent life assistance services, which aid older persons to live longer in their home, have existed for several decades.

In the last few years, due to awareness of the growing distribution of older people, and technological progress, the development of AAL has significantly increased in its importance. Many national and international Non-Governmental Organizations (NGOs) and research projects have been focusing on this topic. As a result, different concepts have entered the market [8][14][15][16]. For instance, through the use of sensory floor mats - which register movements in living areas and react by automatically turning the lights on - the risk of falling can be reduced [17]. Another example of AAL can be found in the combination of personal and technological support offered by the Fraunhofer Institute [8]. By means of summarizing and demand-oriented analysis of sensor data, an individualization of care, as well as nursing services is possible. From a technological perspective it must be noted that most of the described systems are still in their early phase of innovation. Only a few AAL systems are currently marketable [18]. The German Federal Ministry of Education and Research (BMBF) launched the funding program "age-appropriate assistance systems for a healthy and independent life – AAL", which sponsored 18 research projects in the field of AAL with a total amount of € 45 million [19].

Giesecke et al. (2005) have first defined AAL "as the use of AmI [Ambient Intelligence] in everyday life. Assisted means assistance, by technical devices as well as by technical or human services [6]." In 2007, a more elaborate definition of AAL is found in [7]. Hereafter, AAL denotes "living in a smart technology supported environment that reacts sensitively and adaptively to the presence of people and objects and thus provides various services to the human. The aim is to preserve, enlarge and extend the personal freedom and autonomy, by promoting and supporting personal independence [7; translated by the authors]." Although AAL does not explicitly target the elderly and can be implemented in a huge variety of living situations for people with impairments, in practice most of the projects, which carry out research are concerned with the elderly [20]. The definition by Kung and Bart (2010) focuses particularly on enabling older people to experience of a higher quality of life. AAL refers to "intelligent systems that will assist elderly individuals for a better, healthier and safer life in the preferred living environment and covers concepts, products and services that interlink and improve new technologies and the social environment [21]."

AAL cannot be seen as a single technology but as a network of interacting systems or agents, for instance companies from different areas of society. The aim of AAL

is to combine those various agents in one holistic system adapted to diverse customer needs. As seen in [18] four different scopes for application for AAL systems called "health and care", "household and supply", "security and privacy" and "communication and social environment" exist. Due to this diversity, AAL systems should integrate in a modular design and be flexible for the customer's individual needs; lifestyle and health condition [22].

In terms of the German Association for Electrical, Electronic & Information Technologies, AAL is defined as follows: "Assistant systems for the constitution of intelligent environments [aiming] to compensate predominantly age-related functional limitations of different target groups – through technological information and communication support in everyday life [23]." This definition emphasizes the role of information and communication technology in particular.

In contrast to home automation [24] or the smart house [3], AAL is not limited to only life in relation to housing, but extends to all areas of life. AAL focuses on the assistance functions of an adaptive overall system while home automation deals mainly with automation and networking of devices. AAL focuses on maintaining, increasing and extending the user's personal freedom and autonomy. In summary, AAL systems are intended for people with health impairments who require security in their environments and support in communication to prevent loneliness. The present European research focuses on these overall requirements of elderly persons. Since the concept of AAL is concerned with these holistic requirements, the importance of trust in AAL needs to be more understood for permanent usage.

B. The Concept Trust as Influencing Factor for the Usage of AAL

"There are multiple definitions of trust and a single, simple definition is insufficient to capture the essence of the concept [25]." This definition shows the plurality of the concept of trust. As seen in [26] the conception of trust arises in many disciplines like social psychology, philosophy, economics, law, marketing and others. These diverse disciplines also have different basic requirements about trust. The economists have a rational and calculative vision of trust, which contrasts with the attitudinal and ethical view of the philosophers. Social psychology emphasizes the reliability of the word and the fulfillment of obligations [27]. On the other hand, economics perceives trust as an answer to expected future behavior and suggests the usage of hostages to warrant rational behavior [28]. Furthermore, philosophy and social psychology emphasize the personal and interpersonal aspects, while law economics and marketing stress inter organizational trust. The fact that trust depends on additional situational and cultural elements, together with existence of diverse synonyms, highlights the multidimensional view of the concept and demonstrates why there is no uniform definition of the term [26].

Castaldo et al. (2010) used a quantitative approach to illustrate and handle the heterogeneity of trust by means of a content analysis. By application of 36 definitions of the term "trust", a frequency analysis was conducted. The numbers in co-occurrences show that attitude and behavior were used in most of the cases to explain trust [29].

To emphasize the diversity of the construct trust there are added numerous 'trust relationships'. Personal trust, as self-confidence, and interpersonal trust that comprises a human's trust with another human [27][30][31] can be mentioned. [32] put their research focus on close relationships and stated that trust is not present from the beginning. It has to be built up through increasing experience with the other person. Moreover, social trust characterizes trust with a system or an institution [26], while trust in automation denotes a human's trust with a technology or a device [33][34][35].

"Uncertainty, vulnerability and the possibility of avoiding risk or of making a choice based on judgment, are seen as necessary conditions for the existence of trust" [26]. The enhanced uncertainty and complexity that has stimulated the latest interest in trust in various fields of research corresponds with the increased relevance of healthcare and trust in assistance systems and automation in general. Trust in technology induces reliance when the complexness makes a thorough understanding impossible.

Turing (1950) was the first who analyzed trust between humans and machines in an experiment where a human had to differentiate between a human advisor and a computer simulating a human. 95 percent of the participants did not notice the difference and supposed that the advisor was a human. With the 'Turing Effect' the discussion about human trust in information given by automation compared to another human, was born [36].

New and innovative technologies become increasingly complicated and humans cannot manage the full degree of complexity. Humans cannot fully understand the processes behind the automation. They have to rely on automation to use it in an adequate manner. Therefore, trust can be seen as a mediator between humans and automation by guiding reliance: "Trust can be defined as the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [37]. The agent is described as automation or as a person, which cooperates with the surrounding of the person [37].

As seen in different studies, people have the tendency to rely on technology they have trust in and to reject technology they do not trust [33]. When people trust automation, the usage is often influenced positively [38][39]. But also negative examples exist due to inappropriate calibration of user trust. In one notable example, the cruise ship Royal Majesty ran aground because the crew did not realize that the navigation system did not work correctly. The system lost the GPS signal and the alarm did not inform the crew. Although it was obvious to see that the water became too shallow, the disaster was not averted. A subsequent report confirmed that the crew was overly reliant on the automated position display [40]. Another tragic example of distrust towards automation led to an airplane crash where 71 people lost their lives [41]. The collision near Überlingen at Bodensee in 2002 can be attributed to the ignorance towards the Traffic Collision and Alerting System (TCAS). Two airplanes were flying in the same height and the TCAS warned both about the imminent accident. It advised the

Tupolew to ascend and the Boeing to descend but a human air traffic controller was not aware of the other airplane's position. He gave the conflicting advice to the Tupolew to descend. The Tupolew pilots' followed the human's advice and thus the collision was caused. This case represents a typical dilemma of human advisory conflicting with automation advisory. These examples show the importance and impacts of trust towards technology. If trust is not calibrated to the true capacity of the system, users may over rely (misuse) or under rely/ reject (disuse) on the automation [42].

These considerations relating to trust in technology can also have impact in the area of healthcare and AAL. As seen in [43], trust in medical technology is empirically different from trust in other technology. Based on [35], which deal with patients and healthcare providers in obstetric work systems, important implications for trust in healthcare systems and AAL-Technology emerge. The study demonstrates that trust building in medical technology transpires not only in a relationship between doctor and patient or patient and technology. There is a complex network of relationships, which ultimately forms a 'network of trust' in technology use. [44] has already observed a network of trust in supervisory control systems. In addition to the sys-tem she included a system designer, operators, management and society as other actors. Trust as a factor attributed to AAL systems, is also affected by a significant amount of implicit trust in the network around the use of the actual technology. Following the 'Actor Network Theory' [45][46], the reliance on the network located around the AAL system, is equally important for the usage of assistive technology. As an example, for [47] the use of a defibrillator implies not only trust in the product and its functions but also in the network around this product. This network includes the product designer, the organization, which implements the product and the coaches, explaining the technology to the inexperienced users [47]. It follows that distrust in a health care provider can also lead to patients' distrust in medical technology or the hospital per se [48]. Therefore, consideration of the social or work system [49], which encapsulates the technology, is necessary for an understanding of trust. Reference [35] clarified that in the case of complex medical or assistance technology, building trust in automation is more accurately building trust in a work system. Furthermore, during the use of the same system the perspectives of multiple user groups (end user, relatives, and health care provider) vary [35].

In summary, it can be seen that there are a lot of factors, which differ in the formation of trust and, which have to be considered in the development and application of AAL. Due to the importance of the concept of trust it is necessary to develop a working definition as a basis for further research activities in AAL. The following definition based on the above mentioned definitions of AAL [6][21][23] and the definition by [37] in context with automation.

Trust in Ambient Assisted Living (AAL) can be de-fined as the attitude that an assistive technology supports an impaired person within their social environment in an uncertain and vulnerable situation.

AAL also offers holistic support for persons with disabilities, not only to those over the age of 60 years. The combination of human and technical services by modular and customized technology generates various possibilities. Since users will not completely understand the technology and processes of AAL, the attitude trust helps to influence the usage of AAL.

III. RESEARCH FRAMEWORK

A literature review was conducted to explore the relevant scientific approaches in the context of trust in AAL, healthcare assistance systems and other automation. By means of this research method, information about how extensively the issue has been previously addressed in the research can be ascertained. To increase the precision of the literature review in this innovative and fast moving research field, relevant articles were identified by means of computerized search in the online bibliographic databases 'Web of Science' [50] 'PubMed' [51] and 'PsycINFO' [52] starting in November 2010 up to a publication date of January 2011. The three database searches are carried out with filter. In 'Web of Science' key search terms are filtered by topic, in 'PubMed' by MeSH Terms and in 'PsycINFO' by keywords. These three different terminologies represent the generic terms for the search algorithm in the respective database.

For investigation in the three databases, 150 dissimilar search term combinations are performed in each setting. The used key search terms are presented in Table I. The first search requests always contain a term of the categories 'Attributes' and 'Auxiliaries'. At first, the term trust has been set and was queried alternatively with the keywords of the descriptor 'Auxiliaries'. After carrying out these searches, the term reliance was set and also requested with those from the second category. Then, the already carried out 30 search combinations have been linked sequentially to the concepts of the third descriptor 'Population'. By extending the research with these four search terms and consideration of the abbreviations AAL and ATS, ultimately 150 searches per database were performed.

Due to the large number of search combinations and potentially relevant studies, the search results are already reviewed for further availability during the database search. For this, both title and abstract are considered. Afterwards, to identify the relevant full text articles a set of exclusion criteria are selected. For inclusion in the literature review articles had to fulfill the following criteria:

TABLE I. KEY SEARCH TERMS

Attributes	Auxiliaries	Population
Reliance	Ambient Assisted Living/ AAL	Adult
Trust	Assist* System/ Technology	Age*
	Assistive Technology Service/ ATS	Elder*
	Automation	Old*
	Healthcare	
	Intelligent/ Interactive Home	
	Medical Technology	
	Smart Home/ House/ Living	
	Technology	

*Search included stated terms and derivates (e.g., age, aging, aged).

(1) The study described explicitly the connection between trust and automation or assistive technology, whereby trust is seen as an influence factor for the interaction with the system

(2) The article was published in a journal or presented at an international conference

(3) Studies, which were first presented at a conference and afterwards published with identical findings as a journal article, were only taken into consideration with the journal release

(4) The publication was written in English

(5) Due to the database research date, studies are included up until January 2011.

A data form was used to remove the important information for each relevant article. After structuring the articles and integrating the data in the fact sheet, a detailed data analysis was undertaken.

IV. DATA ANALYSIS

The previously described 150 search term combinations in each database initially identified 8,498 potentially relevant articles for the literature review. By means of the structural query, the database 'Web of Science' offered 4,401 publications. The database 'PubMed' yielded 3,855 results and the search requests in 'PsycINFO' provided 242 studies. Owing to the consideration of the above described five exclusion criteria and after analyzing titles and abstracts of the 8,498 studies, 164 publications are used in the next part of the review. In this step, the full text of these 164 articles was reviewed. After analysis of the full text versions, 92 articles were included for the further literature analysis. With regard to the exclusion criteria, a total of 72 of the filtered studies were excluded, leaving 56 percent of the original 164 articles. Fig. 1 gives a numerical overview about the structural sequence of the literature research.

Because of the five exclusion criteria shown above, 72 articles (44 percent) were excluded after the full text review.

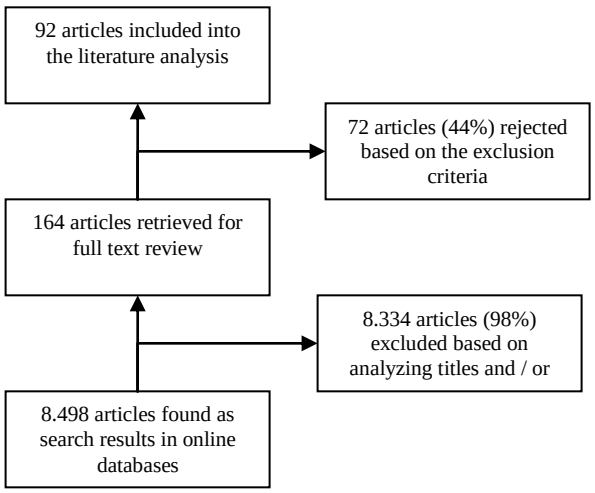


Figure 1. Literature research sequence diagram (authors design).

Most of the studies (48 articles) are not relevant due to the wrong topic focus. 15 of these studies had focused on trust in websites/ online platforms as well as trust in e-commerce applications and are not followed up owing to the exclusion criterion. A further 17 studies are eliminated since they were not published at a conference or in a journal. The last seven excluded articles were published once in a journal and additionally published at a conference with almost identical results. These studies are only considered one time with the more current journal article in our results. Thus, in the end, 92 articles were analyzed in detail in the literature review.

These articles covered the topics trust in automotive [53][54][55][56][56][57], aviation [40][58][59][60][61][62][63][64][65], combat identification [66][67][68][69][70], general design advancement [33][71][72][73][74], supervisory control systems [38][39][75][76][77][78], healthcare and assistance systems [79][80][81][82][83][84] and others [85][86].

As can be seen in Fig. 2, with a total of 18 articles the cluster 'Healthcare and Assistance Systems' has the largest number of relevant studies. This fact can be explained due to the specific key search terms in the descriptor 'Auxiliaries' (e.g., 'Healthcare', 'Assistance/Assistive System/ Technology', or 'Medical Technology') in the first step of the literature search. These articles will be analyzed with special regard in the further course of the study. Firstly, the other six clusters with focus on measuring and analyzing trust in technology will be briefly addressed.

Diverse computerized trials and experiments relating to trust in automation and assistance by means of transportation 'Aviation' (16 articles) or 'Automotive' (11 articles) were performed. In the consideration of the articles in the cluster 'Aviation', the focus is on research in air traffic control (e.g., [62][63][65] and multitask flight simulation [40][59][58]. The background of this field of research is that pilots' trust in alarms or cues within cockpit automation has impact on the usage of autopilot systems. Recent results can be found in [60][61][64].

In the cluster 'Automotive', reliance in automotive

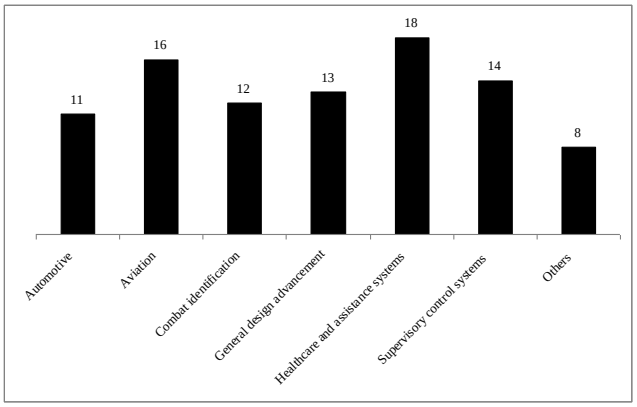


Figure 2. Numerical division of the literature review results by topic (alphabetical order) (authors design).

adaptive cruise control systems [57][87], advanced traveler information systems [55] and particular automotive collision warning systems (e.g., [53][54]) play an important role in research. These computer-assisted experiments aim to analyze trust in different alarm types as false alarm (FA) or unnecessary alarms (UA) in simulated driving situations [56].

Furthermore, in the last decade the military has integrated the factor of trust in 'Combat Identification' in its research projects [66][70]. The literature research included 12 studies on this military subject. The authors in this cluster examined the effects of trust and human responses to automation alerts and false alerts. Participants, who are performing simulated combat tasks, often have to analyze aerial photographs for the presence of enemy targets. Research developments can be seen in [67][68][69].

In addition, 13 articles, among eight former literature reviews, observe trust in 'General Design Advancement' [33][37][73][74]. Jian, Bisantz, and Drury (2000), the only quantitative study in this cluster, developed a trust questionnaire in human-machine interaction, which today is used for measuring trust in various automated systems [71].

Further 14 articles deal with reliance on different 'Supervisory Control Systems'. Monitoring of luggage screening [76], pumping [37][39][78] and central heating systems [77] have been considered in this category. Moreover, Bahner, Hueper, and Manzey (2008) have regarded a process control system and the influence of complacency and automation bias in interacting with a decision aid in this context [75]. Finally, a cluster called 'Others' was created for including all studies, which cannot be integrated in one of the before mentioned sub items. These articles concentrate for example on trust in tele-operation systems [88], automation etiquette [86] or trust in an automated counting and circle estimation task [85].

A. Data analyses of overall results

In the next step, the data sheet with the overall studies has been analyzed (A) and compared with the results from the topic trust in healthcare and assistance systems (B).

(1) Publication date

Between 1987 and 1991 only two studies were published in this context [33][34]. The first experiment of trust in a human-machine supervisory control system was realized by [38]. Whereas, up until 1999, 15 studies were published in total, from 2000 to 2010, 77 articles with regard to trust in technology and assistance systems can be found. Since 2003, every year six studies or more are indicated. In 2008, a maximum of 11 relevant articles are found.

(2) Type of study

In a next step, the distinction between conceptual and empirical/experimental articles is examined. From the overall 92 reviewed studies, 22 consider conceptual and 70 empirical methods for their research. These conceptual articles comprise former summaries and literature reviews (12 articles) as well as articles with the focus on framework, model or questionnaire development (10 articles). The 70 empirical articles can be differentiated into quantitative and qualitative research methods. Since 1987, in total 62 quantitative studies (including experiments, online, postal or paper standardized questionings or a combination of experiment and questioning) were identified. It can be observed, that only five studies include questionnaires only. In contrast, 57 studies used experiments or a combination of experiments and questionnaires for measuring trust. By comparison, eight articles with qualitative methods as qualitative interviews, workshops and focus group interviews were considered.

(3) Participants characteristics

In a next step, the participants' age distribution is considered. In order to receive a better understanding of the participants in experiments or surveys, a clustering into five age groups was conducted. These groups were subdivided into 'participants younger than 30 years' 'participants from 30 to 60 years', as well as 'participants older than 60 years'. Moreover, one age group comprised a combination of younger (< 30 years) as well as older (> 60 years) participants. Further studies performed experiments or interviews without age differentiation.

Regarding the 70 empirical studies, in 22 of the studies or 31 percent, there was no age differentiation declared. In 35 surveys participants were younger than 30 years and in five surveys they were between the ages of 30 and 60 years. In only eight surveys (16 percent of overall) were participants older than 60 years. In five articles the participants exclusively belonged to the age group over 60 years. In three further studies both younger participants (< 30 years) and people over the age of 60 were examined.

Participation rates range from an experiment with six [89] or a qualitative interview with nine participants [90] to a postal survey with 1187 participants [91]. In total, in 43 of the articles (61 percent) less than 50 participants took part in the surveys on trust in automation or assistance systems. In eight studies between 51 and 100 and in 16 studies between 101 and 500 participants were involved. Reference [91] was the only study with more than 500 participants. In two articles there was no participant number specified. Moreover, only three out of the surveys contained a limitation with regard to the gender. One study by [82] questioned 24 women, or rather 24 mothers who had recently given birth. In two other articles only male participants, former pilots [92] and students [93], were surveyed. In 38 surveys both gender were examined and 29 surveys did not make an explicit distinction.

(4) Publication type

Another study detail can be carried out by the differentiation between 'conference vs. journal publication'. Among the 92 examined articles, 18 articles (20 percent) were presented at a corresponding conference and 74 articles published in a journal. The journal with the most publications and major interest in the research of trust and automation was 'Human Factors' with a total of 21 articles (23 percent). The journal 'Ergonomic', with eight relevant articles, the 'International Journal of Industrial Ergonomics', with four and several journals with three studies follow.

B. Data analysis of studies regarding trust in healthcare and assistance systems'

This rising relevance of the concept of trust, which can be found in the different research topics, is also evident by the large number of relevant articles in trust and 'healthcare and assistance systems'. In this field of research interest has been increasing in the last decade.

(1) Publication date

The first published paper relating to trust in healthcare automation was presented in 2002. The conference paper by [86] was the first article that emphasized the factor trust. From this point on until January 2011, 18 articles can be found. These articles deal with reliance on healthcare, medical or household assistance systems. In the years 2003, 2004 and 2006 no publications within this context can be found, whereas since 2007, every year articles are considered. 2010 revealed the largest number of studies in field, with five published. Fig. 3 gives a detailed overview about the annual distribution of the studies in the cluster 'healthcare and assistance systems' in comparison to the other topics. As can be seen, the importance of a conscious handling and perception of the concept of trust in combination with automation and, particularly, healthcare and assistance systems has been increasing in recent years. The first study with regard to trust in automation and human-machine interaction was published in 1987 [33]. In contrast, the first publication regarding trust as a variable for developing healthcare systems for older persons was presented in 2002 [81].

(2) Type of study

Four of the 18 articles in this cluster used conceptual methodologies. Three articles focused on framework or model development [79][81][83] and one study summarized the relevance of training in technology used by tele-home care nurses [84]. Moreover, 14 articles included empirical research–seven used quantitative and seven qualitative methods. The publications with quantitative methodologies are divided into three studies with a combination of questionnaire and experimental design, two studies with exclusive questionnaire surveys and two with experiments. The qualitative research exclusively consists of articles with qualitative interviews.

In comparison, within the other topics quantitative studies dominate with 55 studies. In particular, in new research areas qualitative surveys are utilized to get a detailed understanding of the topic. For this purpose, the focus is set on qualitative interviews, as has occurred in the research area of trust in healthcare and assistance systems. Seven of the overall 18 studies (39 percent) included qualitative interviews with individuals or workshop and focus group discussions. In 2010, four studies used qualitative interviews, which show that researchers are still in the process of developing a detailed understanding. Given that general research on trust in human-machine interaction started in 1987 [33] and to this day ambiguities in this context exist [60][61][94] it is understandable that qualitative interviews are still used in this research area.

(3) Participants characteristics

For the 14 empirical articles an age group differentiation was performed. In three of the studies, participants were younger than 30 years and in one study they are between 30 and 60 years. Moreover, five of the articles consider participants over the age of 60 years. Further two studies consider a combination of younger and older participants, while three surveys give no information about age differentiation. Where the work system is in healthcare and assistance systems such as AAL, the end user is mostly over the age of 60. Therefore, it is of immense relevance that this target group will be considered in the research. Fig. 4 displays the previous study numbers, in which participants over 60 years were involved.

As can be seen, in the other clusters the target group of people over the age of 60 plays only a subordinate role. Only one author has considered elderly persons' trust in a human-decision aid system and compared the results to people younger than 30 years [85].

In contrast, in the healthcare sector researchers have focused more on the age group over 60 years. Of the total of

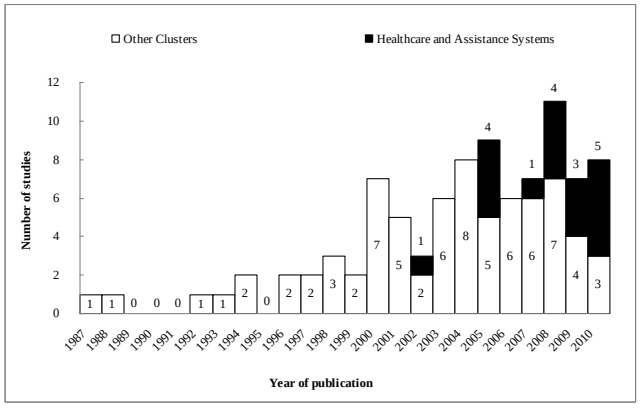


Figure 3. Year distribution of studies in 'healthcare and assistance systems' vs. other topics (authors design).

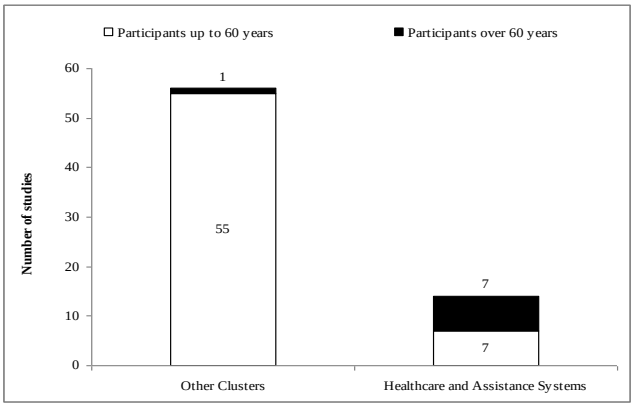


Figure 4. Age differences of study participants in the different clusters (authors design).

eight studies that have dealt with participants over 60 years, seven studies (88 percent) are located in this cluster. Two studies have taken a differentiation of younger and elderly persons into account [80][95]. Moreover, five articles have exclusively focused on people over 60 years [90][96][97][98][99]. In 50 percent of the overall studies, which analyze the factor of trust in healthcare and assistance systems by experiments or surveys, the age group over 60 years is strongly represented.

Concerning the number of participants in the topic 'healthcare and assistance systems', in nine of the studies the participant rate amounted to less than 50 participants. In two studies the participant rate ranged from 51 to 100 persons and three articles took more than 100 participants into account. These articles also include the reference [91] with a postal questionnaire of 1187 people. With reference to the participant rate it can be highlighted that the study with the most participants [91] as well as one of the studies with the least participants (n=9) [90] belong to the topic trust in healthcare and assistance systems.

Regarding the gender distinction within the different methodological designs, 12 articles have regarded both sexes; one article made no differentiation and one study [82] viewed only female participants. This study with solely female participants interviewed 24 women who had recently given birth. They were questioned in a qualitative interview to analyze trust in medical technology and obstetrics work system [82]. For the observation of this complex work system additional interviews with care providers were conducted [35]. Furthermore, it can be said that healthcare and technical support for elderly persons are themes, which concern men and women equally. Therefore, it seems logical that most of the studies deal with both genders.

(4) Publication type

Moreover, among the 18 studies, six articles (33 percent) are presented as conference papers and 12 articles (67 percent) were published in journals. The journal 'Ergonomics' with two publications was the only one, which was represented several times. The author Enid Montague with four research studies since 2009 has taken a pioneering role in context of trust and healthcare technology [35][43][82][83]. Additionally, Coughlin et al. (2007, 2009) and Ho et al. (2005a, 2005b) are listed with two articles [79][80][96][100].

Due to the topical nature of the research field, the distribution of articles presented at conferences and published in journals can be explained. From the overall 18 studies in the healthcare cluster, 12 were published in journals and six studies were presented at conferences. By comparison, from 74 articles within the other topics, 62 were published in journals and 12 studies, thus 16 percent, were presented at conferences.

V. CONCLUSION AND FUTURE WORK

The significant increase of elderly persons due to demographic change and the resulting rise in purchasing power is affecting the development of reliable AAL systems [101]. Since 2005 the European and national sponsoring programs for AAL have steadily increased the relevance of supported living in a home environment, which enlarges and promotes personal independence. Moreover, it is difficult to understand why AAL has had absolutely no consideration in combination with measuring trust in the research literature. The search combinations 'reliance or trust' and 'Ambient Assisted Living/ AAL' yielded no results in the current literature study. There was no study explicitly examining trust in AAL systems. Moreover, the relevance of measuring trust in healthcare technology and assistance systems is not prominent within the research results. It can be seen that the consideration of trust in connection with healthcare, medical technology or assistance systems is still in a nascent stage. A few studies considered trust in intelligent home systems [100], smart home [96], telemedicine systems [98], as well as automation [95] or technology [90] at home. Furthermore, it must be noted, that there is no consistent terminology for assistance systems for elderly persons. No systematic approach and documentation or a uniform technology and understanding exist in research, which complicated measuring trust in this context.

On one hand, these results could imply that the topic has not been viewed as a relevant scholarly topic. On the other hand, due to the increasing number of studies in the last decade, this suggestion seems not to be supported. Analyzing the publication date shows that all relevant articles were composed in this period. It is evident that the research field has gained in importance in the last decade.

Another interesting fact can be found in the different frequency distribution of quantitative and qualitative studies. In the analysis of the type of study it can be highlighted that researchers who are examining trust in healthcare and assistance systems use qualitative as well as quantitative methodologies. The fact that trust in healthcare and assistance systems do not singularly depend on technology but rather on a complex work system [35][47], underlines the relevance of more substantial research into this topic.

Moreover, researchers have recognized that the characteristics of elderly participants have been taken into account. An analysis of trust in this sector can only be realized by the integration of people over the age of 60. Seven articles in the last decade consider older participants' trust in healthcare and home assistance systems. The increasing demand and importance of AAL due to the higher life expectancy and demographic shift clarify a considerable backlog demand in measuring elderly persons' trust in AAL. More research into this age group is required to fill the gap left by the few studies and quantitative results. Finally, it can be surmised that by reason of the novelty of the research of measuring trust as an influencing factor for using healthcare and assistance systems, the exact influence of trust cannot be quantified. Only 18 articles, which cover that topic, were found owing to the literature review. Initial developments reveal that trust in healthcare and medical technology differs from reliance on other technologies [43].

Both, qualitative and quantitative research is required to cope with increasing demands in the coming years. Furthermore, more elderly participants must be taken into account for measuring and conceiving trust in an AAL system. In order to ascertain trust the elderly have in AAL, a

deeper understanding of their needs as well as fears and worries is essential. Additionally, trust of reference persons may have influence in using AAL. For researchers and designers of AAL, recognizing the influencing factor of trust will support the development of marketable solutions.

Due to the knowledge gained by the literature review, further research in the context of elderly persons' trust in AAL will be conducted. Based on the results of the present study, the variables regarding trust in AAL have to be examined in a next step. The investigation conforms the various influence factors on trust in AAL and beyond the connection to the usage intention. For this, a scenario-based questionnaire survey and additional experiments will be performed. The experiments include mock-ups of AAL technology on tablet PCs. Different scenarios will be conducted by older test persons. The impact of personal and technical assistance within AAL will be examined and afterwards reliability of AAL technology manipulated.

LIMITATIONS

The systematic review had to contend with some limitations in the research process. First, the selection of online databases should be considered. Literature for trust in automation and healthcare can be seen as an interdisciplinary field. Therefore, three bibliographic databases were used: 'Web of Science' comprising of interdisciplinary content across 256 disciplines; the database 'PubMed' focusing on healthcare content; and 'PsycINFO', psychological literature. Due to this selection, articles, which are not integrated in these databases, are excluded for the review. Second, the information provided in the articles is very heterogenic. Some include a specific description about the experimental design, while other studies fail to provide detailed information. Third, due to the fact that only English language articles were included in the review, a distorted picture is drawn, as the studies focus on English-speaking authors. Fourth and finally, the studies included in the literature review were screened up until January 2011. Thus, articles, which were published afterwards, are not considered for this systematic review.

REFERENCES

[1] F. Steinke, T. Fritsch, and L. Silbermann, "A Systematic Review of Trust in Automation and Assistance Systems for Older Persons' Overall Requirements," eTELEMED 2012 The Fourth International Conference on eHealth, Telemedicine, and Social Medicine, Valencia, 2012, pp. 155-163.

[2] W. Paulus, J. Hilbert, and W. Potratz, "ICT for Housing," in Information and Communication Technologies for Active Ageing. Opportunities and Challenges for the European Union, N. Malanowski and M. Cabrera, Eds. Amsterdam, 2009.

[3] G. Leopold, "Will 'Smart House' Provide Shelter for High-Tech Firms?," Electronics, vol. 58(26), 1985, pp. 45-46.

[4] M. Spring, "Home smart home," Building, vol. 30, 1986.

[5] B. Philips, and H. Zhao, "Predictors of assistive technology abandonment," Assistive Technology, vol. 5, 1993, pp. 36-45.

[6] S. Giesecke et al., "AAL - Ambient Assisted Living: Country Report Germany, White Paper, 2005.

[7] T. Becks, J. Dehm, and B. Eberhardt, "Ambient Assisted Living. Neue "intelligente" Assistenzsysteme für Prävention, Homecare und Pflege," Frankfurt am Main, 2007.

[8] Fraunhofer, „Zuhause Daheim: Das Projekt JUTTA," 2011, retrieved on 12-11-2012, retrieved from http://www.inhaus.fraunhofer.de/Geschaeftsfelder/Health-und-Care/jutta.jsp.

[9] UNDESA, "World Population Ageing 2009," Department of Economic and Social Affairs: Population Division, New York, 2010.

[10] European Union. "The 2012 Ageing Report: Underlying Assumptions and Projection Methodologies," European Economy April 2011, retrieved on 11-12-2012, retrieved from http://ec.europa.eu/economy_finance/publications/european_economy/2011/pdf/ee-2011-4_en.pdf.

[11] World Health Organization (WHO), "Health statistics and health information systems: Definition of an older or elderly person," 2011, retrieved on 11-12-2012, retrieved from http://www.who.int/healthinfo/survey/ageingdefnolder/en/index.html.

[12] I. Stuart-Hamilton, "The psychology of ageing: an introduction," 2n ed. London: Jessica Kingsley Publishers, 1994.

[13] J. Hilbert, K. Scharfenorth, and J. Haberle, "Vom Virtuellen Altenheim zu TESS inkontakt. Erfahrungen aus einem Entwicklungs- und Erprobungsprojekt für mehr Lebensqualität im Alter," IAT, Ed., Jahrbuch 1998/1999, 1999, pp. 132-143.

[14] BMBF, "Assistenzsysteme im Dienste des älteren Menschen. Steckbriefe der ausgewählten Projekte in der BMBF-Fördermaßnahme „Altersgerechte Assistenzsysteme für ein gesundes und unabhängiges Leben – AAL", 2010, retrieved on 11-12-2012, retrievd from http://www.vde.com/de/Technik/AAL/Steckbriefe/Documents/sens@home.pdf.

[15] W. Heusinger, "Das intelligente Haus- Entwicklung und Bedeutung für die Lebensqualität," Frankfurt am Main, 2005.

[16] S. Solaimani, H. Bouwman, and M. de Reuver, "Smart Home: Aligning Business Models and Providers Processes; A case survey," Proceedings of the 21st Australian Conference on Information Systems Aligning Business Models and Providers Processes, Brisbane, 2010.

[17] C. Sälzer, "Besser wohnen mit Technik?," Das AALmagazin, vol. 2, 2010, pp. 12-16, retrieved on 11-12-2012, retrieved from http://www.smartliving-gmbh.de/downloads/Besser%20Wohnen%20mit%20Technik.pdf.

[18] P. Georgieff, "Ambient Assisted Living: Marktpotenziale IT-unterstützter Pflege für ein selbstbestimmtes Altern," Stuttgart: MFG Stiftung Baden-Württemberg, 2008.

[19] Bundesministerium für Bildung und Forschung (BMBF), "Assistenzsysteme im Dienste des älteren Menschen – Steckbriefe für ausgewählte Projekte in der BMBF-Fördermaßnahme ‚Altersgerechte Assistenzsysteme für ein gesundes und unabhängiges Leben – AAL'," 2009, retrieved on 11-12-2012, retrievd from http://www.bmbf.de/pubRD/projekte_aal_2012.pdf.

[20] R. G. Heinze, and C. Ley, "Abschlussbericht des Forschnungsprojektes Vernetztes Wohnen. Ausbreitung, Akzeptanz und nachhaltige Geschäftsmodelle," T-Mobile Deutschland GmbH (Bonn). Bochum, 2009, retrieved on 11-12-2012, retrieved from http://www.sowi.rub.de/mam/content/heinze/heinze/abschlussbericht_vernetzteswohnen.pdf.

[21] A. Kung, and B. Jean-Bart, "Making AAL Platforms a Reality," Proceedings of the AMI 10 workshop (AmI-10). Ambient Intelligence. Lecture Notes in Computer Science, 2010, vol. 6439, 2010, pp. 187-196.

[22] R. Wichert, "Challenges and limitations of intelligent ambient assisted living environments," in: Ambient intelligence. (de

Ruyter, B. et al., eds.), Proceedings of the First international joint conference, AmI 2010, Málaga, Spain, November 10-12, 2010, Springer, pp. 304–309.

[23] M. Gersch, R. Lindert, and M. Hewing, "AAL-business models: Different prospects for the successful implementation of innovative services in the primary and secondary healthcare market," Proceedings of the AALIANCE European Conference on AAL, Malaga, 11-12 March 2010, retrieved on 12-11-2012, retrieved from http://www.aaliance.eu/public/oldstuff/aaliance-conference-1/papers-and-posters/8_2_fu-berlin.

[24] I. Miles, "<Shift> <Control>,...<Home>? A Response to Robins and Cornford," Futures, vol. 22(8), 1990, pp. 880-885.

[25] D. C. P. Holland, and P. A. G. Lockett, "Business trust and the formation of virtual organizations," Proceedings of the 31 Annual Hawaii International Conference on System Sciences, volume IV, 1998, pp. 602-609.

[26] K. Blomqvist, "The many faces of trust," Scandinavian Journal of Management, vol. 13(3), 1997, pp. 271-286.

[27] J. B. Rotter, "A new scale for the measurement of interpersonal trust," Journal of Personality, vol. 35, 1967, pp. 651–665.

[28] N. G. Noorderhaven, "Trust and inter-firm relations," Proceedings of the 1992 EAEPE Conference, 1992.

[29] S. Castaldo, K. Premazzi, and F. Zerbini, "The Meaning(s) of Trust. A Content Analysis on the Diverse Conceptualizations of Trust in Scholarly Research on Business Relationships," Journal of Business Ethics, vol. 96, Number 4 (2010), pp. 657-668.

[30] M. Deutch, "Trust and suspicion," Journal of Conflict Resolution, vol. 2, 1958, pp. 265–279.

[31] M. Deutch, "The effect of motivational orientation upon trust and suspicion," Human Relations, vol. 13, 1960, pp. 123-139.

[32] J. K. Rempel, J. G. Holmes, and M. P. Zanna, "Trust in close relationships," Journal of Personality and Social Psychology, vol. 49, 1985, pp. 95-112.

[33] B. M. Muir, "Trust between humans and machines, and the design of decision aids," International Journal of Man Machine Studies, vol. 27, 1987, pp. 527–539.

[34] T. B. Sheridan, "Trustworthiness of command and control systems," in 3. IFAC/IFIP/IEA/IFORS Conference on Analysis; Design and Evaluation of Man-Machine Systems, Oulu, Finland, 1988, pp. 14-16.

[35] E. N. H. Montague, W. W. Winchester III, and B. M. Kleiner, "Trust in medical technology by patients and healthcare providers in obstetric work systems," Behaviour & Information Technology, vol. 29(5), 2010, pp. 541-554.

[36] A. M. Turing, "Computing machinery and intelligence. Mind," vol. 59, 1950, pp. 433-460.

[37] J. D. Lee, and K. A. See, "Trust in automation: Designing for appropriate reliance," Human Factors, vol. 46(1), 2004, pp. 50–80.

[38] J. Lee, and N. Moray, "Trust, control strategies and allocation of function in human-machine systems," Ergonomics, vol. 35, 1992, pp. 1243–1270.

[39] J. D. Lee, and N. Moray, "Trust, self-confidence, and operators' adaptation to automation," International Journal of Human-Computer Studies, vol. 40, 1994, pp. 153–184.

[40] R. Parasuraman, and C. Miller, "Trust and etiquette in high-criticality automated systems," Communications of the Association for Computing Machinery, vol. 47(4), 2004, pp. 51–55.

[41] BFU, "Untersuchungsbericht AX001-1-2/02. Edited by Bundesstelle für Fluguntersuchung," 2004, retrieved on 11-12-2012, retrieved from http://www.bfu-web.de/nn_41670/DE/Publikationen/Untersuchungsberichte/2002/Bericht__02__AX001-1-2,templateId=raw,property=publicationFile.pdf/Bericht_02_AX001-1-2.pdf.

[42] R. Parasuraman, and V. Riley, "Humans and automation: use, misuse, disuse, abuse," Human Factors, vol. 39(2), 1997, pp. 230–253.

[43] E. N. H. Montague, B. M. Kleiner, and W. W. Winchester III, "Empirically understanding trust in medical technology," International Journal of Industrial Ergonomics, 2009, doi:10.1016/j.ergon.2009.01.004.

[44] B. M. Muir, "Trust in automation: I. Theoretical issues in the study of trust and human intervention in automated systems," Ergonomics, vol. 37(11), 1994, pp. 1905–1922.

[45] M. Callon, "Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St Brieuc Bay," in "Power, Action and Belief: a new Sociology of Knowledge? Sociological Review Monograph," J. Law, Ed., London, 1986, pp. 196-233.

[46] B. Latour, "Science in Action: How to Follow Scientists and Engineers Through Society," Milton Keynes, 1987.

[47] S. Timmons, R. Harrison-Paul, and B. Crosbie, "How do lay people come to trust the Automatic External Defibrillator?," Health, Risk & Society, vol. 10(3), 2008, pp. 207-220.

[48] E. N. H. Montague, and B. M. Kleiner, "Using variance analysis to detect mismatches in role expectations in patient physician interactions in obstetric work systems," Paper presented at the 2009 International Ergonomics Association, Beijing, China, 2009.

[49] H. W. Hendrick, and B. M. Kleiner, "Macroergonomics: An introduction to work system design," Santa Monica, CA: Human Factors and Ergonomics Society, 2001.

[50] ThomsonReuters, "Web of Science," 2011, retrieved on 06-27-2011, retrieved from http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/.

[51] PubMed, 2011, retrieved on 12-11-2012, retrieved from http://www.ncbi.nlm.nih.gov/pubmed.

[52] PsycINFO, 2011, retrieved on 12-11-2012, retrieved from http://www.apa.org/pubs/databases/psycinfo/index.aspx.

[53] G. Abe, and J. Richardson, "Alarm timing, trust and driver expectation for forward collision warning systems," Applied Ergonomics, vol. 37, 2006, pp. 577-586.

[54] J. S. Chugh, and J. K. Caird, "In-Vehicle Train Warnings (ITW): The Effect of Reliability and Failure Type on Driver Perception Response Time and Trust," in Human Factors and Ergonomics Society 43rd Annual Meeting, Houston, TX, 1999.

[55] B. H. Kantowitz, R. J. Hanowski, and S. C. Kantowitz, "Driver acceptance of unreliable traffic information in familiar and unfamiliar settings," Human Factors, vol. 39, 1997, pp. 164–176.

[56] M. N. Lees, and J. D. Lee, "The influence of distraction and driving context on driver response to imperfect collision warning systems," Ergonomics, vol. 50(8), 2007, pp. 1264–1286.

[57] N. A. Stanton, and M. S. Young, "Driver behaviour with adaptive cruise control," Ergonomics, vol. 48(10), 2005, pp. 1294- 1313.

[58] N. Bagheri, and G. Jamieson, "The impact of context-related reliability on automation failure detection and scanning behavior systems," Proceedings of the 2004 IEEE International Conference, Man and Cybernetics, vol. 1 (10–13), 2004, pp. 212-217.

[59] J. P. Bliss, and M. C. Dunn, "Behavioural implications of alarm mistrust as a function of task workload," Ergonomics, vol. 43(9), 2000, pp. 1283-1300.

[60] J. S. Hughes, S. Rice, D. Trafimow, and K. D. Clayton, "The automated cockpit: A comparison of attitudes towards human

and automated pilots," in Transportation Research, vol. F12(5), 2009, pp. 428-439.

[61] A. Keller, and S. Rice, "System-Wide versus Component-Specific Trust Using Multiple Aids," The Journal of General Psychology, vol. 137(1), 2010, pp. 114–128.

[62] A. J. Masalonis et al., "Air traffic controller trust in a conflict probe during Free Flights," Paper presented at the 42nd Annual Meeting of the Human Factors and Ergonomics Society, Santa Monica, CA, 1998.

[63] J. M. McGuirl, and N. B. Sarter, "Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information," Human Factors, vol. 48(4), 2006, pp. 656–665.

[64] S. Rice, and K. Geels, "Using System-Wide Trust Theory to Make Predictions About Dependence on Four Diagnostic Aids," The Journal of General Psychology, vol. 137(4), 2010, pp. 362-375.

[65] C. D. Wickens et al., "False alerts in air traffic control conflict alerting system: Is there a "cry wolf" effect?," Human Factors, vol. 51(4), 2009, pp. 446-462.

[66] M. T. Dzindolet, L. G. Pierce, H. P. Beck, and L. A. Dawe, "The perceived utility of human and automated aids in a visual detection task," Human Factors, vol. 44, 2002, pp. 79–94.

[67] S. Rice, "Examining single- and multiple-process theories of trust in automation," The Journal of general psychology, vol. 136(3), 2009, pp. 303-319.

[68] S. Rice, K. D. Clayton, A. Wells, and D. Keller, "Manipulating Trust Behaviors in a Combat Identification Task," Workshop Human Factors Issues in Combat Identification, 2008.

[69] L. Wang, G. A. Jamieson, and J. G. Hollands, "Trust and Reliance on an Automated Combat Identification System," Human Factors, vol. 51(3), 2009, pp. 281-291.

[70] M. Yeh, and C. Wickens, "Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration," Human Factors, vol. 43(3), 2001, pp. 355-365.

[71] J. Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," International Journal of Cognitive Ergonomics, vol. 1(4), 2000, pp. 53-71.

[72] J. D. Lee, and K. A. See, "Trust in automation: Designing for appropriate reliance," Human Factors, vol. 46(1), 2004, pp. 50–80.

[73] P. Madhavan, and D. A. Wiegmann, "Similarities and differences between human–human and human–automation trust: an integrative review," Theoretical Issues in Ergonomics Science, vol. 8(4), 2007, pp. 277–301.

[74] R. Parasuraman, and C. D. Wickens, "Humans: Still Vital After All These Years of Automation," Human Factors, vol. 50(3), 2008, pp. 511–520.

[75] J. E. Bahner, A.-D. Hueper, and D. Manzey, "Misuse of automated decision aids: Complacency, automation bias and the impact of training experience," International Journal of Human-Computer Studies, vol. 66, 2008, pp. 688–699.

[76] S. M. Merritt, and D. R. Ilgen, "Not all trust is created equal: dispositional and history-based trust in human-automation interactions," Human Factors, vol. 50(2), 2008, pp. 194-210.

[77] N. Moray, T. Inagaki, and M. Itoh, "Adaptive automation, trust, and self-confidence in fault management of time-critical tasks," Journal of experimental psychology, vol. 6(1), 2000, pp. 44-58.

[78] D. A. Wiegmann, A. Rich, and H. Zhang, "Automated diagnostic aids: The effects of aid reliability on users' trust and reliance," Theoretical Issues in Ergonomics Science, vol. 2(4), 2001, pp. 352–367.

[79] G. Ho, L. M. Kiff, T. Plocher, and K. Z. Haigh, "A model of trust and reliance of automation technology for older users," Papers of the AAAI Fall Symposium "Caring Machines: AI in Eldercare," Menlo Park, CA, 2005, pp. 45-50.

[80] G. Ho, D. Wheatley, and C. T. Scialfa, "Age differences in trust and reliance of a medication management system," Interacting with Computers, vol. 17, 2005, pp. 690–710.

[81] C. A. Miller, K. Haigh, and W. Dewing, "First, cause no harm: Issues in building safe, reliable and trustworthy elder care systems," Proceedings of the AAAI-02 Workshop "Automation as Caregiver", 2002.

[82] E. N. H. Montague, "Patient source of learning about health technologies and ratings of trust in technologies used in their care," Ergonomics, vol. 53(11), 2010, pp. 1302–1310.

[83] E. N. H. Montague, "Validation of a trust in medical technology instrument," Applied Ergonomics, 2010, doi:10.1016/j.apergo.2010.01.009

[84] K. Shea, and J. A. Effken, "Enhancing Patients' Trust in the Virtual Home Healthcare Nurse," Computers, Informatics, Nursing, vol. 26(3), 2008, pp. 135–141.

[85] N. Ezer, A. D. Fisk, and W. A. Rogers, "Age-Related Differences in Reliance Behavior Attributable to Costs Within a Human-Decision Aid System," Human Factors, vol. 50(6), 2008, pp. 853-863.

[86] C. A. Miller, "Trust in adaptive automation: the role of etiquette in tuning trust via analogic and affective methods," Proceedings of the First International Conference on Augmented Cognition, Las Vegas, NV, 2005.

[87] B. Rajaonah, F. Anceaux, and F. Vienne, "Trust and the use of adaptive cruise control: a study of a cut-in situation," Cognition, Technology & Work, vol. 8(2), 2006, pp. 146-155.

[88] I. Dassonville, D. Jolly, and A. M. Desodt, "Trust between Man and Machine in a Teleoperation System," Reliability Engineering and System Safety, vol. 53(3), 1996, pp. 319-325.

[89] X. Jiang et al., "Measurement of human trust in a hybrid inspection system based on signal detection theory measures," in International Journal of Industrial Ergonomics, vol. 34, 2004, pp. 407-419.

[90] B.-M. Wälivaara, S. Andersson, and K. Axelsson, "Views on technology among people in need of health care at home," International Journal of Circumpolar Health, vol. 68(2), 2009, pp. 158-169.

[91] M. Calnan, D. Montaner, and R. Horne, "How acceptable are innovative health-care technologies? A survey of public beliefs and attitudes in England and Wales," Social Science & Medicine, vol. 60, 2005, pp. 1937–1948.

[92] Y. J. Tenney, W. H. Rogers, and R. W. Pew, "Pilot opinions on cockpit automation issues," International Journal of Aviation Psychology, vol. 8, 1998, pp. 103–120.

[93] B. M. Muir, and N. Moray, "Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation," Ergonomics, vol. 39(3), 1996, pp. 429-460.

[94] R. Parasuraman, and D. Manzey, "Complacency and bias in human use of automation: An attentional integration," Human Factors, vol. 52(3), 2010, pp. 381-410.

[95] J. Sanchez, G. Calcaterra, and Q. Q. Tran, "Automation in the home: The development of an appropriate system representation and its effects on reliance," Proceedings of the Human Factors and Ergonomics Society 49th annual meeting, 2005.

[96] J. F. Coughlin, L. A. D'Ambrosio, B. Reimer, and M. R. Pratt, "Older Adult Perceptions of Smart Home Technologies: Implications for Research, Policy & Market Innovations in Healthcare," Proceedings of the Engineering in Medicine & Biology Annual Conference IEEE, Lyon, 2007.

[97] C. Harrefors, K. Axelsson, and S. Sävenstedt, ”Using assistive technology services at differing levels of care: healthy older couples’ perceptions,” Journal of Advanced Nursing, vol. 66(7), 2010, pp. 1523–1532.

[98] Y. Jasemian, “Elderly comfort and compliance to modern telemedicine system at home,” Proceedings of the Second International Conference on Pervasive Computing Technologies for Healthcare, PervasiveHealth 2008, Tampere, 2008.

[99] M.-L. Jung, and K. Loria, “Acceptance of Swedish e-health services,” in Journal of Multidisciplinary Healthcare, vol. 3, 2010, pp. 55–63.

[100] J. F. Coughlin, J. Lau, L. Ambrosio, and B. Reimer, “Adult children's perceptions of intelligent home systems in the care of elderly parents,” Proceedings of the 3rd International Convention on Rehabilitation Engineering & Assistive Technology, 2009.

[101] tns emnid, “Wohnwünsche im Alter,” Grafikreport, Januar 2011, retrieved on 11-12-2012, retrieved from http://www.bfw-bund.de/uploads/media/Emnid_Wohnw%C3%BCnsche_im_Alter_-_Pressemappe.pdf.

Pairing Prolog and the Web in a Normalization and Denormalization Tool

Lule Ahmedi
Computer Engineering
University of Prishtina
Prishtina, Kosova
lule.ahmedi@uni-pr.edu

Naxhije Jakupi
Computer Science
South East European University
Tetovo, FYR Macedonia
naxhijejakupi@yahoo.com

Edmond Jajaga
Information Technoogies
State University of Tetovo
Tetovo, FYR Macedonia
e.jajaga@unite.edu.mk

Besmir Sejdiu
Computer Engineering
University of Prishtina
Prishtina, Kosova
besmir.sejdiu@gmail.com

Abstract—**The concept of database normalization emerges as very important since the need of redundant-free storages. Its hard understanding by students requires a flexible learning environment and an intelligent behavior by the machine. Our tool NORMALDB integrates both of them in a PHP and Prolog implemented platform providing a complex architecture internally and user friendly interface. Students will effectively learn by working with their own defined examples and at the same time can consult the theory behind each normalization step. On the other side, database teachers can track the tool usage of their students, by getting the information of their actions. This paper provides a detailed description of NORMALDB pointing out the communication between HTML pages and Prolog predicates and spread the challenges faced during the application deployment. Instead of normalization, our tool is the first in the series to support the teaching of denormalization procedure.**

Keywords-E-learning systems; database normalization and denormalization; logic programming; prolog server; web programming;

I. INTRODUCTION

Learning technologies is one of the fields that highlighted the potential of the web for education [1]. Easy access, location and time independent resources, unlimited design space, flexibility, and a wide range of functionalities of the web prompted the development of a new branch of learning named e-Learning [2]. The advantages of the web were by default inherited by e-Learning technologies built upon them, enabling thus more sophisticated teaching and learning environments. That way, the utilization of e-Learning technologies by universities has led them to transform from didactic teaching methods to flexible and independent learning. Passing to this level of authomatizm requires introducing the latest programming technologies, through which the machines will behave more like human. Algorithms that simulate thought processes and reasoning that produce behavior similar to humans belongs to a special field called Artificial Intelligence (AI). These software portions of applications are known as intelligent agents, which can learn from the interaction of the human and the machine.

The AI based e-Learning applications have the potential of producing realistic environments in which students can perform learning. In this simulated environment the student interacts with the intelligent agents, which in turn perceive changes and take appropriate actions. Such a feature is very important in the domain of e-Learning, especially in subject-based e-Learning. These applications are an example of providing support for even more complex subjects to learn [3], as is database normalization and denormalization, which is a subject-to-learn through our e-Learning tool we will here introduce.

Database design is the art and science of improving the structure of database relations that are most suited to represent a small portion of the world called the "universe of discourse" [4]. The relational schema that results at the end of the design phase must consist of normalized relations accordant with the semantics of given entities and their integrity constraints (ICs), avoiding at the same time as more as possible data manipulation anomalies. Hence, normalization is very important in practice, but also crucial to get familiar with, for every student studying databases. Unfortunately, this subject is often dry and troublesome to learn, making it not well received by students.

To ease learning of database normalization and denormalization, we developed NORMALDB, a web-based e-Learning tool. It is designated to provide theoretical background on the subject (normalization and denormalization of database schema): it explains stepwise every single detail of the process as a whole. It also provides an interactive interface of learning the subject, driven by the student's given examples. The content organization of the tool is a pure reflection of how the teacher organizes the subject.

The rest of the paper is organized as follows. Section II describes the current state-of-the-art on e-Learning tools for database normalization. A brief review on the main concepts of database normalization and denormalization is given in Section III, Section IV introduces the NORMALDB's system architectural layers and its features. Further, Section V provides the main challenges faced during the system design, whereas Section VI describes the learning methodology using NORMALDB. Section VII includes deployment of NORMALDB in real accessible platforms for testing purposes. Finally, the paper ends with the conclusion and future works section.

II. RELATED WORK

There are already few tools that cover executing normalization, like JMathNorm [5], NORMIT [6], RDBNorma [7] and Micro [8]. To the best of our knowledge, NORMALDB is the first which supports the denormalization step.

JMathNorm [5] is an interactive tool for relational database (RDB) normalization implemented by using Mathematica. The tool includes normalization till Boyce-Codd Normal Form (BCNF) and calculates closure and cover of functional dependencies (FDs). The design approach of our system is similar to this tool. JMathNorm integrates Mathematica for writing normalization modules and Java language for programming the user interface, while we use Prolog and PHP, respectively. Our advantage over JMathNorm is the advantage of using web-based systems over the standalone ones in terms of software distribution, compatibility issues and immediate software updates.

Another tool dealing with the database normalization, NORMIT [6], is a web-enabled Intelligent Tutoring Systems, which is also the first in the series of constraint-based tutors developed at ICTG (University of Canterbury). The emphasis of this system is on problem solving thus complementing the students' class work. Its domain knowledge is modeled using Constraint-Based Modeling (CBM), which enables a student to work on with normalization problems as long as he/she reaches a state known to be true pre-defined by 53 constraints. With NORMALDB our emphasis was on establishing an e-Learning tool which shall not only be complement to the class work, but also to be an independent database normalization "teacher". Additionally, our tool is more complete w. r. t. the inclusion of the normalization concepts like finding minimal cover, making the decomposition lossless and projecting FDs.

Micro [8] and RDBNorma [7] are tools that employ linked lists to model the relation entity along with its FDs. Micro uses two linked lists to store the relation, the first one stores all the relation attributes while the other one stores FDs. RDBNorma focuses on system performance in terms of space and time consumption by using only one linked list to represent a relation along with FDs holding on it. In our tool the relation schema and FDs are modeled using two separate Prolog lists. Our emphasis during the system design was on providing a more user friendly GUI and a flexible learning environment, while authors of RDBNorma does not deal with user interface issues.

Generally, NORMALDB is unique in that it smoothly integrates two entirely diverse paradigms, namely:

- at the internal level, an intelligent layer based on logic rules in Prolog that implements the normalization and denormalization of a given database, whereas,
- at the external level, the user may friendly and stepwise interact with the tool through a common Web interface, kept thereby not concerned with the complexity of the tool at its internal level.

Moreover, using NORMALDB, students may step-by-step experience the whole life-cycle of database relations up to their normalized forms, or in a reverse process of denormalization, which means roll backing relations into their original form whenever deemed necessary for the sake of efficiency of join operations. Driven by examples, each step is in addition accompanied with comprehensive explanation of the theories applied in that given step. Navigation through content blocks across individual steps / subtopics is another strong reason to leverage NORMALDB. The level of interactivity, the ease of use, its logic-base of rules and the Web interface make NORMALDB unique among existing tools, which support normalization.

III. DATABASE NORMALIZATION AND DENORMALIZATION

It has been estimated that more than 80% of all computer programs are database-oriented. This is easy to believe since databases allow the applications to meet all their requirements for storing, manipulating and displaying data [9] at once.

For years now, the relational data model remains the most used data model in databases. The central data description construct in this model is a relation, which can be thought of as a set of records. The description of a data in terms of data model is called a schema. In relational model, the schema for the relation specifies its name, the name of each field, and the type of each field. User requirements may in addition result into certain ICs within the schema. ICs may in turn cause redundancy-related problems like: redundant storage, update anomalies, insertion anomalies, and deletion anomalies. Special group of ICs that plays the major role in the schema refinement are called FDs [10].

Together with the input schema, FDs provide the initial information from which it is produced normalized relations i.e., anomalies-free relations. Namely, FDs holding over a relation influence a relation to be split in two or more relations. This technique is known as relation decomposition. A decomposition of a relation schema R consists of replacing the relation schema by two (or more) relation schemas that each contain a subset of the attributes of R and together include all the attributes in R [10]. Whether to decompose a relation or not, it depends on the desired level of redundancy. Additionally, in order to not losing any information when performing relation decomposition we need to be aware of two issues:

- Lossless-join decomposition, which enables to recover any instance of the decomposed relation from the corresponding instances of the smaller relations.
- Dependency-preserving decomposition, which prevents from expensive joins of derived relations by enforcing original relation's FDs on each of the derived relations.

A. *Normalization*

Following the FDs that hold over a relation, one may understand what redundancy problems, if any, might arise from the current schema. To provide such guidance, several normal forms (NFs) have been [10] introduced in terms of FDs as follows:

- 1NF – First Normal Form: A relation R is in first normal form if and only if all underlying domains contain atomic values only.
- 2NF – Second Normal Form: A relation R is in second normal form if and only if it is in 1NF and

every nonkey attribute is fully dependent on the primary key.
- 3NF – Third Normal Form: A relation R is in third normal form if and only if it is in 2NF and every nonkey attribute is nontransitively dependent on the primary key.
- BCNF – A relation R is in BCNF if and only if every determinant is a candidate key.
- 4NF and 5NF are rarely achieved, and hence not implemented in our tool at this stage.

The procedure itself of transforming a relation, given its FDs, into any of the abovementioned NFs is known as normalization, and it:
- leaves the relation unchanged if it already satisfies the NF sought after, or
- decomposes the relation in two or more smaller relations, i.e., relations with less number of columns, each satisfying the NF sought after.

Regarding the normalization theory it is necessary to mention the concept of attribute closure, since every normalization algorithm, including 3NF and BCNF algorithms, within the Prolog system for normalization [4] uses it. Closure of a set of attributes X with respect to a set of FDs as the set of all attributes A such that X → A can be deduced by Armstrong's axioms [4, 10].

B. Denormalization

Normalization of a relational schema with the given set of FDs results into a relational schema, which is free of redundancy-derived anomalies, but might yet suffer from eventual performance-derived problems. To address that kind of problems, a reverse process to normalization, known as denormalization, has been introduced. Denormalization is the process of adding columns to the table to reduce joins in favor of performance, and is considered only if the integrity of data is not seriously compromised [11].

The two most common types of denormalization are:
- **Two Entities in a One-to-One Relationship:** The tables for these entities could be implemented as a single table, thus avoiding frequent joins required by certain applications.
- **Two Entities in a One-to-many Relationship:** Sometimes logical design results in very simple tables with very few attributes, where the primary key is a foreign key in another table you want to join with. In such cases, when a query wants data from both tables, it may be more efficient to implement them as individually named columns as an extension of the parent entity (table).

The procedure of denormalization need to follow these steps:

First of all there is a need to make sure that normalization process was done by correctly applying NF rules. Then will be selected the dominant queries and updates based on the criteria such as high frequency of execution, high volume of data accessed, response time constraints or explicit high priority. This analysis will result on definition of the tables that require extra columns, when appropriate to reduce the number of joins required for dominant queries. After the recomposition of the tables with extra colums there is a need to consider also the data integrity due to denormalization [9].

Denormalizing databases is a critical issue because of the important trade-offs between system performance, ease of use, and data integrity [12]. Thus, a database designer should have a good reason when deciding to perform denormalization.

IV. THE ARCHITECTURE OF NORMALDB

As stated in [11], students find it difficult to understand the concept of FDs and normalize data in order to obtain smaller well-structured relations. NORMALDB is a web-based e-Learning tool that we developed to aid students understand and experience the most complex tasks of database design, i.e., normalization and denormalization. The organization of NORMALDB resembles the way how a teacher schedules his / her class while teaching normalization and denormalization. Further, the ability to explore every single step / subtopic of normalization by running student's own examples and breaking them down to elementary details makes NORMALDB far more advantageous vs. traditional in-class teaching of the subject.

In the following subsections is given a description of NORMALDB to reflect its implementation in two layers, the logical and the interface layer.

A. Logical Layer

1) Normalization: At the data tier and business logic layer of NORMALDB, we adopted the Ceri and Gottlob's script [4] implemented in Prolog. The Prolog programming language is known for its contributions to problem solving in AI [13]. A common integrated framework for describing both data structures ("facts") and algorithms ("rules"), and the facilitated interaction with the code through the "trial and error" interface are few among several advantages readily provided by the Ceri and Gottlob code due to the logical representation of normalization in Prolog [4].

The adopted Prolog script [4] of normalization consists of the following:
- the facts, which provide data about relations (the relation name, attributes, and FDs), and
- the rules, which relate facts, and implement all algorithms throughout normalization.

For example, according to [4], a relation schema `rel` with the set `[a, b, c]` of attributes, and with set of FDs `[a → b, b → c]` is represented with the following facts at our logical layer:

```
schema(rel, [a, b, c]).
fd(rel, [a], [b]).
fd(rel, [b], [c]).
```

For each normalization step, there is a rule or a set of rules that may be invoked in any order. This way a user may

observe results incrementally by executing certain rules step by step over the input base of facts. Some of the normalization rules provided in the script are as follows [4]:

- `findonekey(REL, K)` - Determines one key K of relation REL
- `assertallkeys(REL)` - Determines and asserts all keys of REL
- `findmincover(REL)` - Finds a minimal cover of the FDs defined for REL
- `thirdnf(REL)` - Decomposes REL into 3NF
- `haslj(REL)` - Tests for losslessness of the decomposition of REL
- `makelj(REL)` - Makes the decomposition of REL lossless
- `projectfds(REL, RELl)` - Projects FDs holding for the relation REL to the relation RELl
- `isinbcnf(REL)` - Tests whether REL is in BCNF
- `bcnf(REL)` - Decomposes REL into BCNF
- `formminimize(REL)` - Minimizes the decomposition of REL

2) Denormalization: Denormalization also supported in NORMALDB is a rather intuitive task driven primarily by queries, which are frequently invoked in a database and involve expensive joins.

The implementation in NORMALDB of denormalization extends the existing Prolog knowledge base of normalization with the following:

- rules for the direct re-composition of tables, and
- a parser written in Definite Clause Grammar (DCG) notation of Prolog, which is able to read queries against the database and reason upon them to infer, which tables need re-composition in favor of performance.

Regarding the feature of re-composition of relations we have extended the Prolog code with a pair of rules.

The first one, rule `recompose`, gets the two relations schema to be merged (row 2 and 3), generates a new name for the newly created relation (row 4), ensures there is not any other relation schema in the base of facts with the one it is trying to create by deleting any `NEWREL` instance (row 5), merges two schemas in a new one `NEWSCHEMA` (row 6) and finally calls the rule `recomposerel`, which will build the new relation NEWREL with its schema `NEWSCHEMA`.

```
1.    recompose(REL, REL1, REL2)-->
2.    {schema(REL1, SCH1),
3.    schema(REL2, SCH2),
4.    makerecomposedname(REL1, NEWREL),
5.    retractall(schema(NEWREL, K)),
6.    append(SCH1, SCH2, NEWSCHEMA)},
7.    html(div(\recomposerel(REL,
      NEWREL, NEWSCHEMA))).
```

The second one, rule `recomposerel`, asserts the newly created relation schema to the base of facts, asserts the decomposition information between the original relation and the newly re-composed relation, projects the FDs from the original relation schema to the recomposed relation, minimizes the original relation i.e., if eventually the recomposed relation schema is a subset of another existing relation schema it is discarded from the database, and finally an output information is published in a HTML page created by HTML generators.

Another improvement to the Prolog script was the Prolog DCG parser. Semantic grammar is an engineering technique for constructing natural language understanding systems [14]. The denormalization algorithm of the tool is written with this notation of Prolog. SQL queries given by the user will be read using DCG grammar and then processing them will result in a statistical table of joins between relations. A detailed description follows in the next subsection.

3) Rendering Prolog results into HTML. In addition to denormalization, few more modifications to the Prolog script of Ceri and Gottlob [4] were applied to enable the integration with the web-based module, i.e., adding new rules and facts for rendering Prolog results into the web page. Most of new rules are HTML generators, which convert the adopted Prolog script to the Prolog server [15], namely they generate HTML tags, which hold the results retrieved from Prolog predicates.

These generators are built-up using DCG notation rules. The idea behind this is to translate a Prolog term i.e., predicate into an HTML document. HTML generators are predicates from html_write.pl library. This library contains a wide range of DCG rules that enables easy generation of HTML pages.

The DCG non-terminal `html/3 (html(:Spec))` is the main predicate of this library. It translates the specification for an HTML page into a list of atoms that can be written to a stream using `print-html/1`. The expansion rules of this predicate may be extended by defining the multi-file DCG `html_write: expand/1`. Spec is either a single specification or a list of single specifications.

Nested with this predicate can be used another predicates that are build up as per request, which contain some Prolog rules and then the result from this rule is injected within any HTML tag by using any `html` predicate.

B. Interface Layer

Next we discuss the interface layer of NORMALDB, which is mainly developed in PHP language.

Next we discuss the interface layer of NormalDB, which is mainly developed in PHP language. The HTML tag rendering is provided through PHP scripts, whereas JavaScript, especially its libraries jQuery [16] and jQuery UI [17] help make the NORMALDB interface simple and easy to navigate through.

Figure 1 shows the organization of NORMALDB from user perspective. The colored boxes represent web pages, whereas the grey box represents the knowledge base of the application supplied by the Prolog server, which runs whenever a normalization step is called. Connection to the knowledge base does not require any extra procedure from the user, except for a user to be signed in. Simply, the given

connection hyperlink makes a request to the Prolog server, which in turn displays the result.

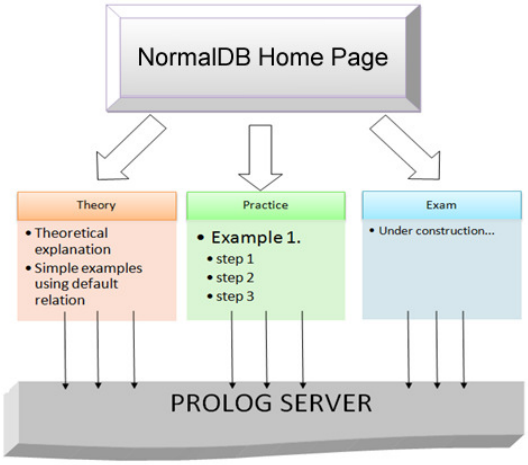


Figure 1. Organization of the interface layer in NormalDB.

The interconnection between topics and examples is done via hyperlinks, while results are represented in accordance to the actual web page template. Depending on the content, the result appears within a dedicated tag of the page, or in a message box that shows after clicking the link. The JavaScript language and its libraries jQuery and jQuery UI are employed for user-friendly purposes of the interface.
jQuery tabs and message boxes helped us prevent the overload of the interface. Also the management of the Document Object Model (DOM) is carried out with jQuery routines. This is applied when the student gives initial relation schema attributes and FDs. The remove() and appendTo() methods are employed to add/remove attributes and FDs.

The whole interface of the tool (colored boxes in Figure 1) is organized in three main parts:

Theoretical part: Consists of theoretical explanations of the topics covering the normalization phase of the database design. Each topic explanation appears as a hyperlink, and may further contain links, which illustrate the application of the given theory to a given example. That example is referred to as a default example in our tool, and is designed to serve the demonstration of each of the theories over a given set of (default) relations and their FDs.

Practical part: This is the core part of NORMALDB, and is aimed to serve exercises. The user (e.g., a student) shall input the relation name, attributes, as well as FDs of the relation that he / she wants to examine in terms of normalization and denormalization, and the tool will then start exploring the topics upon the given relation.

Exam part (Self-assessment part): This part is planned for future work. It is designed to provide a testing environment where users (e.g., students) may themselves examine their knowledge gained in the field concerning both exercises and the theory.

V. IMPLEMENTATION

The development environment in building NORMALDB consisted of the scripting language PHP, JavaScript, and the logic programming language Prolog.

The logic of the NORMALDB relays in the server side of the application. The client side shares just the functionalities given by JavaScript codes.

The server side of the application is comprised of two distinct servers. Apache server is employed for processing PHP scripts and generating HTML pages for the client, while Prolog server provides the knowledge base of the system. The later one is accessed from the client side through links generated from the PHP scripts.

Figure 2 describes the used architecture in NORMALDB with the pursued workflow of the functionalities. In the following subsections we will explain the challenges that appeared during the system design and implementation..

A. *Prolog Server*

Prolog is an excellent tool for representing and manipulating data written in formal languages as well as natural languages. Its safe semantics and automatic memory management make it a prime candidate for programming robust Web services [15].

There are two views on deploying Prolog for Web related tasks [15]:

- **Prolog like an embedded component**: Prolog acts as an embedded component in a general Web processing environment. In this view it is a component that can be part of any of the layers of the popular three-tier architecture for Web application. Components generally exchange XML if used as part of the backend or middleware services and HTML if used in the presentation layer.
- **Stand-alone application:** HTTP allows Prolog to be deployed in any part of the service architecture, including the realization of complete Web applications in one or more Prolog processes.

The later one is the view that is used in our project. The reason why this separation is done is based on the way how Prolog reacts to the inter-platform communication. There are web applications, which require the execution of the Prolog logic only once. Thus, the best suited solution is to follow the application of Prolog as embedded component. Other applications require the Prolog knowledge base to remain active during the user session. In this case the second view is recommended.

In NORMALDB the knowledge base changes several times through normalization steps. Because of this we followed the second approach by employing Prolog server features for handling the communication between HTML and Prolog.

Figure 2 describes how the communication flows between servers and client browser.

Whenever an HTTP request for PHP processing is sent, the Apache server responds with its results.

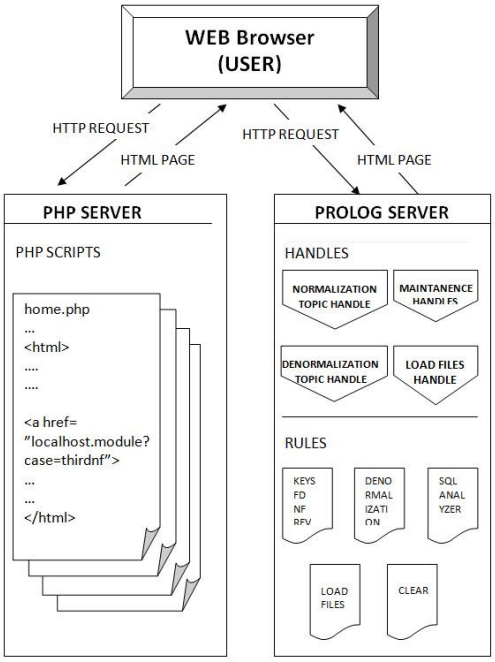


Figure 2. Workflow and request chains of NormalDB tool.

On the other side, when an HTTP request dedicated for Prolog processing is sent to the Prolog server, the later one responds with returning its results. The requests to the servers aredone from the user via clicking to the links appeard in HTML pages.

This way NORMALDB responds to the user requests by exchanging data on two servers at the same time, which illustrates the flexibility of the tool.

B. *Preserving the state*

The process of normalization and denormalization flows over a step by step evaluation, which requires keeping and following the active state of the script execution. This happens because of the modifications that are done to the knowledge base after evaluating a particular step. Traversing through the Prolog script to find the predicates that will implicate the required rule, results in asserting new facts that affects the next step. For example, if a user wants to test whether a particular decomposition has the lossless join property, a clause of the form `haslj(rel)` is searched in the base of facts. This kind of facts is provided after the execution of the 3NF algorithm.

This way it was necessary to have the knowledge base updated with the information required by users' future request. Hence, it was needed somehow the Prolog application to be active as long as a user session is active.

When addressing this concern, we ended up with three alternative solutions, each applying distinct techniques originating from different fields.

One solution was to run the Prolog script using the `system()`, `exec()`, and `shell_exec()` built-in PHP functions. These functions are easy to implement, but the troubles arise after running a required predicate since the script then closes up such that the newly asserted predicates cannot be saved.

Another solution was to use a relational database at the backend of the application, which will track every inference deduced by the Prolog script, and accordingly modify respective tables in the database. Yet for the sake of the simplicity of the tool, and to avoid difficulties that might appear while tracking Prolog inferences, this solution was omitted.

Finally, we experienced the use of Prolog server [15] as the most appropriate solution for surpassing this problem, which will be discussed in detail in the following subsections.

C. *Communication with Prolog server*

The Prolog logic programming language supports a number of libraries for accessing data on HTTP (Hypertext Transfer Protocol) servers, as well as for providing HTTP server capabilities from SWI-Prolog. Both server and client are modular libraries. The server can be operated from the `Unix inetd super-daemon`, as well as a stand-alone server that runs on all platforms supported by SWI-Prolog [17].

In order to use these libraries, certain modifications to the actual normalization script in Prolog of Ceri and Gottlob were required. Thereby, the mere logic of the script is kept unmodified, extending it with built-in predicates to deal with HTTP requests to configure the HTML pages.

In a traditional web environment user actions are captured by clicking the links. The links on the pages are formulated based on the desired action. For example, if the user lands in a page for finding the minimal cover of the FDs then the link containing the appropriate handle together with the input variables is constructed by using PHP string manipulation functions and is rendered to the anchor tag. The input variables in most of the cases consist of the name of the relation instance in which the minimal cover will be computed (`bank`) and the name of the predicate in the Prolog script for performing the calculation (`findmincover`).

```
<a href="http://localhost:5000
       /module?name=findmincover
               &relation=bank">
```

Library `library(http/http_dispatch)`, which is defined in the beginning of the script, dispatches the request from HTTP request, where it gets the location of the predicate that will serve the page from the URL and find the required handler.

Request handlers are built-in predicates that handle the HTTP requests made from instantiated HTML pages using hyperlinks. When an HTTP request arrives at the server, then Prolog starts traversing through the predicate tree for finding the handle that matches the path came with the HTTP request. The required path is noted as first parameter of the

predicate. The second parameter defines the main predicate that handles the handler. As the third parameter of this handler is the list of the options related with the handler. A code that creates a common handler is written below:

```
:- http_handler(root(module),home,[]).
```

After calling a handler, a thread is employed to search for a predicate supported by the handler, which is defined in the code.

```
home(Request) :-
http_parameters(Request,
    [name(Name,[length >= 1]),
    relation(Rel,[length >= 2])]),
reply_html_page(title('Example'),

    [\html_requires(css('style.css')),
       \case(Name, Rel)]).
```

The HTTP request can include variables, which can be extracted from the request and inherited to other predicates that build up the rule. The base predicate that is retrieved from the handler renders the HTML page with enclosed html, title and body tags, by the following rule:

```
home(Request):-http_parameters(Request,
[name(Name,[length >= 2]),
relation(Rel,[length >= 2])]),
reply_html_page(title('Example'),
[\html_requires(css('style.css')),
\case(Name, Rel)]).
```

The body part of the new arranged HTML page is filled with other HTML tags that are derived from the next HTML generators used from other predicates.

Weaving of the HTML tags with the appropriate results is done through HTML tag generators included in the special defined rules, which are called with DCG notation of Prolog. These rules get the results inferred from the base predicates and render them into HTML tags that can be read from every web browser.

The formulated HTML reply of the Prolog server is injected into the HTML page rend from the Apache server within a `<iframes>` tag.

A sample Prolog rule that makes the rendering of the result of 3NF normalization in a special `div` tag looks like follows:

```
case(thirdnf,Rel)-->
    {cleandecomp(Rel)},
    html(div([\step1_html(Rel),
          \step2_html(Rel),
          \step3_html(Rel),
          \step4_html(Rel),
          \step5_html(Rel)])).
```

Some of the links are processed at runtime during the execution of the PHP page, while others are simple links and are invoked when a user clicks them.

D. Request processing inside Prolog server

After the appropriate handle is activated, the traversing inside the Prolog server rules and predicates starts in hierarchical manner.

The built-in rule `http_parameters` starts the procedure with extracting the parameters that are required from the following predicates. This predicate fetch and type-check parameters transparently for both GET and POST requests and converts them to atoms.

Parameters that were catch from the HTTP request, after converting to atoms, are passed to the main predicates that will render the HTML page.

```
reply_html_page(title('Example'),
  [\html_requires(css('style.css')),
  \case(Name, Rel)]).
```

The body content of the output pages is build using HTML tag generators written in DCG notation. The main tag generator that is used here is `html/1` predicate. Within this generator can be used other tag generators, like div, table, h2 and so on.

```
html(div([table(class(stats),
   [\theader([ 'Name',
               'Attribute',
               'Relation Key'])
   | \rowssimple(L)])])).
```

E. Preserving Consistency

The consistency of the Prolog server will be intimidated if the same predicate is queried two or more times without rolling back to the actual state of the server. Multiple queries are posed in case the user clicks the same link multiple times. To avoid unwanted results from the repeated clicks, there are specific rules asserted in Prolog server that take care to clear the server from the previous results.

Additionally, using the same relation name within the same knowledge base by different users will raise the problem of interference among results of different users. To avoid this, we used the unique session number functionality of PHP. The opened index page automatically creates a new session with unique number, which will uniquely identify the relation schema used by that session, concretely by one user. The task of Prolog server in this case is to create a copy of the default relation schema and to rename it with a combination of its original name and the session number of the user.

In this way, normalization rules consider every user session as unique, enabling thus every user work with exercises separately at the same time on the same server.

F. SQL Parser for denormalization

In the denormalization phase, there is no rule that indicates how to recompose resulted tables from

normalization. Instead, this phase is more intuitive and based on the queries that could be used by the end-user or data warehouse transactions..

To give this logic to our tool, it is required to have the module that will "understand" the queries posed by the user and to calculate how many times each table is joined with another one.

This knowledge is represented in Prolog using DCG notation, which is specific for natural language processing tasks.

The script starts with the rule:

```
sql(s(S,F,K)) -->  select(S),
                   from(F),
                   where(K).
```

which seperates the given query in three parts and continues with analyzing the content of each part.

In the body part of these parsing rules are placed predicates that count the joined relations after detecting the names of tables that are joined within query. This rationale has been implemented with the following predicate:

```
joining(T1, T2) :-
          (joined(T1,T2,N)),M is N+1,
          retractall(joined(T1,T2,N)),
          assertz(joined(T1,T2,M)).
```

VI. CASE STUDY: LEARNING NORMALIZATION IN NORMALDB

Accessing NORMALDB through a URL will initially open its homepage as is common for web applications (Figure 3). As mentioned earlier in Section IV, the tool is organized in three individual modules: theoretical, practical, and the testing module.

The *theoretical module* (opens when clicking the green box in Figure 3) loads a sample example, which illustrates all steps of normalization one after another running in the Prolog server. The relational schema used in the example is named *rel* and is given a set of predefined FDs.

The normalization steps or subtopics are each provided on a left menu in the page (Figure 4). Changing along the menu links, which represent subtopics, does not affect the Prolog server unless a button involving the example is clicked.

Such buttons contain links to the Prolog server, invoking thus the corresponding predicate to run for the given subtopic.

The left menu contains also the denormalization module of the tool. This page, beside the theoretical explanation of the topic, includes a form where the user may textually input queries supposed to be executed over time against the derived database schema. These queries are then analyzed with a SQL (Structured Query Language) analyzer script written in Prolog, which yields the statistics about which tables need to be joined. This way, the user may decide which tables need to be recomposed.

Figure 3. Home Page

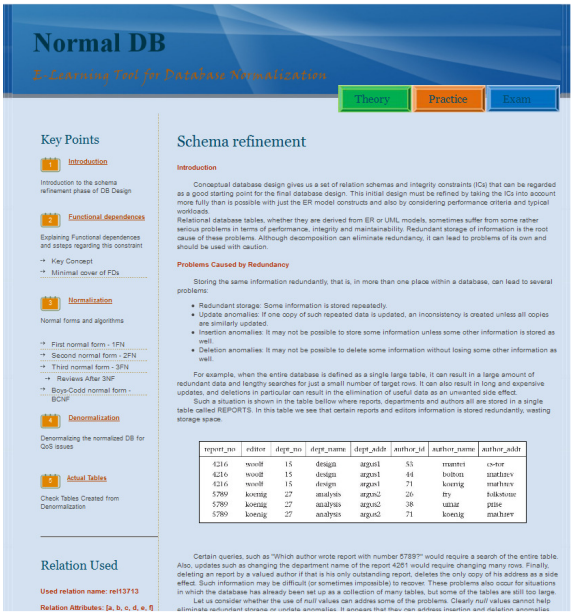


Figure 4. Theoretical part

The result representation in the application is done by using jQuery [16] controls. These controls are activated by clicking the buttons, which send appropriate requests to Prolog server.

In theoretical part, these functionalities are placed in the end of each introduced section, whereas in the practical part these are the main functionalities of the pages.

Navigation to the *practical module* is possible from any page (by clicking the orange box placed at the top of the page), not just while residing at the NORMALDB homepage. Once opened, this module (see the Web page in Figure 5) will first retract all facts belonging to the sample example of the theoretical part, and afterwards eventually load a new relational schema entered interactively by the user through a Web form. If the user provides also the relation attribute types, then the tool will be able to generate a ready-to-deploy SQL script consisting of procedures for creating the database and its tables.

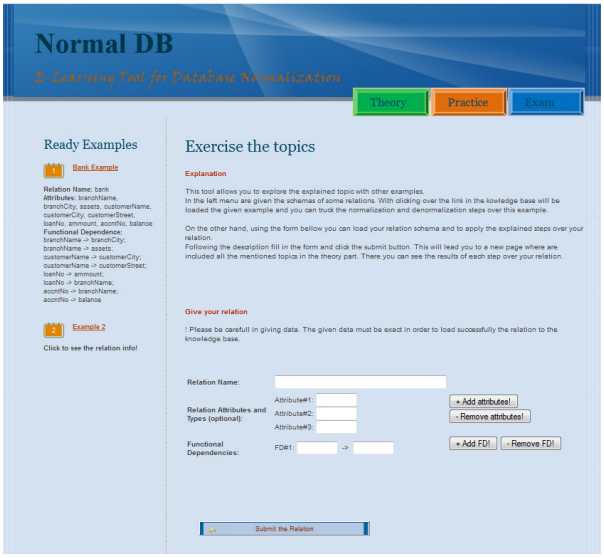


Figure 5. Practicall Part: Form for giving new relation

The relation schema, its attributes and FDs entered through the Web form are processed by PHP string manipulation built-in functions, and are written as new facts in a separate Prolog file. For instance, considering the example of Section III, let us assume that the user enters the same input data, i.e., the `rel` relation schema with attributes `[a, b, c]` and a set `[a → b, b → c]` of FDs through the Web form of the practice page as depicted in Figure 5.

In the left menu of this page, are listed different ready examples. By clicking to one of them, the relation that is described will be loaded and then the user can continue exercising with the given topics. Figure 6 shows new loaded ready-given exercise named "bank"

Exercising panel for both, ready-given examples and examples given by users, is the same page. This shows also the dynamicity of NORMALDB.

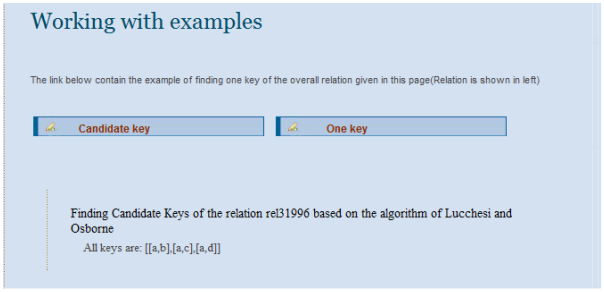
Figure 6. Practicall Part: Ready given example

After uploading the new relation (given by user or ready given example), all links that will be used for further explanation of the topic will use this relation as reference and will apply their rules to this relation.

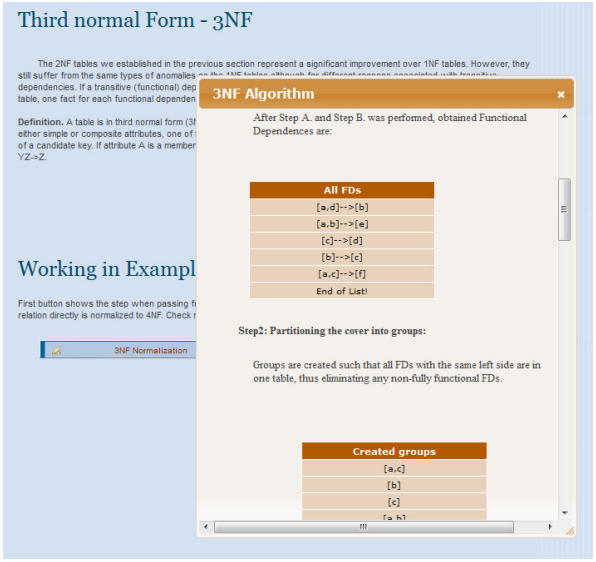
Whenever the user wants to change the relation instance he may move to the last tab and give new relation instance data. When the new relation will be uploaded, the data about the old one will be deleted and removed from the knowledge base.

The upper mentioned initial facts will be stored in the server in a separate Prolog file and will be eventually consulted during the lifetime of the user session (Figure 5).

After uploading the initial data the first thing a user can do is to find all relation keys by clicking the *Assert Key* button, which actually triggers the evaluation of the rule `assertallkeys(bank)` at the Prolog server. As illustrated in Figure 7a the result of our example will be "`All keys are [[branchcame, branchcity, assets, customername, customerstreet, loanno, accntno]]`", rendered at the bottom of the page.



a) Result rendered within the page



b) Result appears in new message box

Figure 7. Result rendering in practical part

The other display format of results is by using jQuery dialog box. This format is used when the resulting display contains more than one Prolog rule execution. Figure 7b displays the 3NF decomposition of our running example.

As one may experience NORMALDB has met the requirement of simplicity in its interface and usage. The organization of the page, its rich set of features and the simple layout contribute altogether towards bringing closer to students the normalization theory, which has otherwise proved to be troublesome to capture in a traditional teaching classroom.

VII. APPLICATION DEPLOYMENT

In order for users to be able to use this tool, we have published it on the Internet and have established an authentication module for accessing the tool. Because the platform of NORMALDB includes the scripting language PHP, JavaScript and the logic programming language Prolog, the PHP scripting files have been published on Apache Web Server port 80, which was installed on a Linux platform and the Prolog server was activated within SWI-Prolog server for handling Prolog requests. The HTTP Prolog server library consists of two mandatory parts and one optional part [15]. The first mandatory one deals with connection management, while the second mandatory one implements a generic wrapper for decoding the HTTP request calling for user code to handle it and encode the answer. Dealing with hosting of the Prolog server, three different approaches were considered:

- One is to run it on a dedicated machine [15] on port 80, the standard HTTP port. The machine may be a running virtual one. The (virtual) machine approach isolates security threads and allows for using a standard port.
- The server can also be hosted on a non-standard port such as 8000, or 8080. However, using non-standard ports may experience issues from intermediate proxy- and/or firewall policies.
- Another approach is to use Apache reverse proxies. This causes the main web-server to relay requests below a given URL location to our Prolog server.

In our case the Prolog server, has been hosted on a dedicated machine on port 80, a standard HTTP port, on Windows platform. Installation of the SWI-Prolog on a Windows system contains the following important features [15]:

- A folder called pl containing the executables, libraries, etc. of the system. No files are installed outside this folder.
- A program swipl-win.exe, providing a window for interaction with Prolog. Another program, swipl.exe, is a version of SWI-Prolog that runs in a DOS-box.
- The file-extension .pl is associated with the program swipl-win.exe. Opening a .pl file will cause swipl-win.exe to start, will switch the directory to the one in which the file-to-open resides and load this file. Additionally, other Prolog files may be consulted within the SWI-Prolog.

A. *Security, Authentication, and Event Logging*

Web security is a set of procedures, practices, and technologies for assuring a reliable, predictable operation on web servers, web browsers, other programs that communicate with web-servers, and the surrounding Internet infrastructure. Unfortunately, the sheer scale and complexity of the Web makes the problem of web security dramatically [18] more complex than the problem of Internet security in general. Today's web security problem has three primary facts:

- Securing the web-server and the data that is on it;
- Securing information that travels between the web-server and the user;
- Securing the end user's computer and other devices that people use to access the Internet;

NORMALDB authentication module offers three different user groups: students, instructors and administrators, where each of the user groups has different permissions. A one's access is enabled through traditional login data with username and password, where the password is encrypted with SHA-1 (Secure Hash Algorithm).

In each page of NORMALDB authentication control is performed, and in the case there is unauthorized attempt for access, the session automatically will be redirected to the login page. For security reasons, during the logging process some of the user's data are recorded, like: Username, Hostname – the computer name from which the user has tried to log in, Ipaddress – user's IP address, AttemptStatus – status of the attempt, which may take value Successful or Failed, and EntryDate – date and time when the attempt was made.

NormalDB also provides the feature of event logging where every user's action gets evidence. The logging event includes the following data: Username – the user who approaches the tool, SessionId – a unique ID that identifies each session to the application, EntryDate – date and time when the attempt was made, Actions – describes the actions taken by the user e.g., New Example means the user has practiced a new normalization example, and Details – contains information about the input relational schema and the input FDs. Figure 8 shows a view of event logging.

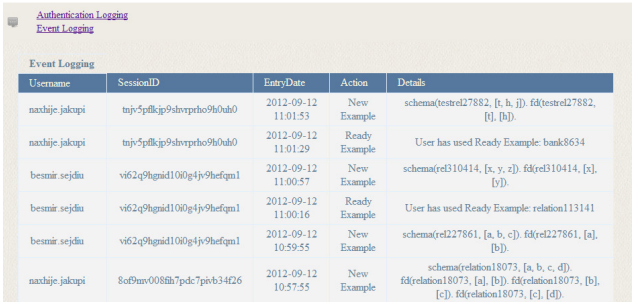


Figure 8. Event logging in NormalDB

VIII. CONCLUSION AND FUTURE WORK

This paper introduced NORMALDB, an e-Learning tool, which offers a user-friendly environment for learning and experimenting the normalization phase of the database design. Simple interface with an internal complex system and a wide range of features including experiments with self-chosen examples are some of the main advantageous features of NORMALDB. For the purposes of friendlier UI and efficient client-server interaction several jQuery and jQuery UI routines were applied. Additionally, some Prolog rules were added to the Prolog script for database normalization [4], mainly for supporting denormalization feature of our tool.

During the development phase different challenges were encountered. The main one was to enable an efficient communication between PHP and Prolog. Among several alternatives considered, the interaction through a Prolog server [15] proved to be the most appropriate solution for our application since it already provides a number of Prolog libraries for accessing data on HTTP. Syntactically, the communication between HTML and Prolog predicates has been carried out through DCG notation. The HTML rendering is produced through HTML tag generators.

The tool was deployed in two separate servers, namely one for handling PHP requests – Apache server, and another for handling Prolog reasoning – Prolog server. While deploying Prolog server we eventually encountered difficulties, which were solved with deploying it into a dedicated Windows-based hosting server.

By placing NORMALDB in real environments, it was possible to check and test the communication between Apache and Prolog server. We plan to evaluate NORMALDB by placing it for use by database students among our future plans around the NORMALDB.

NORMALDB is further planned to expand its capabilities of a typical e-Learning tool by incorporating a test module where the student may test himself / herself by solving a given normalization problem, and then comparing his / her solution with the one generated by the tool. Moreover, the support for higher NFs is one of our future plans, besides adding new visual effects that will enable graphical interpretation of relations' decomposition. Adding higher NFs is easy to implement, since the interface is flexible for adding new features and as stated in [4] the knowledge base can be expanded.

REFERENCES

[1] L. Ahmedi, N. Jakupi, and E. Jajaga, "*NORMALDB – A logic-based interactive e-Learning tool for database normalization and denormalization*", The Fourth Intl. Conf. on Mobile, Hybrid, and On-line Learning (eL&mL), Jan. 30- Feb. 4, 2012, pp. 44-50.

[2] Namahn, "E-*learning: A research note by Namahn*", www.namahn.com/resources/documents/note-e-learning.pdf, 22.11.2011.

[3] J. O. Uhomoibhi, "*Implementing e-learning in Northern Ireland: prospects and challenges*", Campus-Wide Information Systems, vol. 23 no.1, 2006, pp. 4-14, doi:10.1108/10650740610639697.

[4] S. Ceri and G. Gottlob, "*Normalization of relations and PROLOG*", Commun. ACM , vol. 29 no. 6, 1986, pp. 524-544, doi:10.1145/5948.5952.

[5] A. Yazici and Z. Karakaya, "*JMathNorm: A Database Normalization Tool Using Mathematica*", ICCS (2), Lecture Notes in Computer Science , vol. 4488, 2007, pp. 186-193, doi:10.1007/978-3-540-72586-2_27.

[6] A. Mitrovic, "*NORMIT: A Web-Enabled Tutor for Database Normalization*", Intl. Conf. on Computers in Education (ICCE), Auckland, New Zealand, 3-6 Dec. 2002, pp. 1276-1280, doi:10.1109/CIE.2002.1186210.

[7] P. S. Dhabe, Y. V. Dongare, and S. V. Deshmukh, "*RDBNorma: - A semi-automated tool for relational database schema normalization up to third normal form*", International Journal of Database Management Systems (IJDMS), vol. 3, no.1, Feb. 2011.

[8] Du H and Wery L, "*Micro: A normalization tool for relational database designers*" Journal of Network and Computer application, vol. 22, Oct. 1999, pp. 215-232, doi: 10.1006/jnca.1999.0096.

[9] R. Stephens, "*Beginning Database Design Solutions*" Wiley Publishing, 2008.

[10] R. Ramakrishnan and J. Gehrke, "*Database Management Systems",* 2nd ed., McGraw-Hill, 2002.

[11] S. S. Lightstone, T. J. Teorey, and T. Nadeau, "*Physical Database Design: the database professional's guide to exploiting indexes, views, storage, and more*" 4th ed., Morgan Kaufmann, 2007.

[12] G. L. Sanders and S. K. Shin, "*Denormalization effects on performance of RDBMS*", Proceedings of the 34th Annual Hawaii International Conference on System Sciences (HICSS-34), Jan. 2001.

[13] G. F. Luger, "*Artificial Intelligence: Structures and Strategies for Complex Problem Solving",* 6th ed., Addison Wesley, 2008.

[14] R. R. Burton and J. S. Brown, "*Semantic grammar: A technique for constructing Natural Language interfaces to instructional systems*", 1977.

[15] J. Wielemaker, SWI-Prolog HTTP support. SWI-Prolog's home page:
http://www.swi-prolog.org/pldoc/package/http.html, 22.11.2011.

[16] jQuery Documentation. Sept. 2009. The jQuery Project: http://jquery.org/, 22.11.2011.

[17] jQuery User Inerface Library. Sept. 2009. The jQuery UI library: http://jqueryui.com/, 22.11.2011.

[18] S. Garfinkel, "*Web Security, Privacy and Commerce",* 2nd ed., O'Reilly, November 2001.

[19] K. Hsiang-Jui and T. Hui-Lien, "*A web-based tool to enhance teaching/learning database normalization*", Proc. of the 2006 Southern Association for Information Systems Conf., 2006, pp. 251-258.

[20] D. Akehurst, B. Bordbar, P. Rodgers, and N. Dalgliesh, "*Automatic Normalisation via Metamodelling*", ASE 2002 Workshop on Declarative Meta Programming to Support Software Development, Sept. 2002.

[21] S. Ram, "*Teaching data normalisation: Traditional classroom methods versus online visual methods - A literature review*", Proc. of the 21st Annual Conf. of the National Advisory Committee on Computing Qualifications, Auckland, New Zeland, 2008, pp. 327-330.

Impacts of Culture on Gesture Based Interfaces:
A Case Study on Anglo-Celtics and Latin Americans

Manolya Kavakli and Karime Nasser
Department of Computing
Macquarie University
Sydney, Australia
Manolya.kavakli@mq.edu.au, Karimenasser@hotmail.com

Abstract- **This paper investigates the impacts of culture in gesture-based interface design. The aim is to identify basic cultural differences in using hand gestures between two cultures: Anglo-Celtics and Latin Americans. We videotaped and analyzed 10 participants in two groups, while describing two chairs by using their speech and hand gestures. We investigated the frequency, occurrence, and the type of gestures used by the two cultures, as well as the words most frequently used by the participants. Our findings state that there are cultural differences in hand gestures during the description of the objects. This may have implications for the development of gesture-based multimodal interfaces. Anglo-Celtics coming from a low context culture, describe objects using a larger number of gestures as well as a larger vocabulary in a longer period. On the other hand, Latin Americans coming from a high context culture, use smaller number of gestures more frequently in a shorter period. The differences in frequency of gestures may have an impact on the adoption of new technologies as well as adaptation to them. We also found that as the complexity of a task increases, so does the number and type of gestures used. Our conclusion is that the gesture vocabulary of a multimodal interface will be affected not only by the complexity of the task being performed, but also by the cultural background and the language skills of the user.**

Keywords- Gesture recognition; Cultural difference; Gesture based interface design; Gesture segmentation; Speech coding.

I. INTRODUCTION

This study builds on and extends our previous research on the "Influence of culture in multimodal interfaces" [1] and as its predecessor, aims at defining the variances in gesture behavior from one culture to another. Culture can be defined as the shared patterns of behaviors and interactions, cognitive constructs, and affective understanding that are learned through a process of socialization. These shared patterns identify the members of a cultural group while also distinguishing those of another group [2]. Our goal in this paper is to identify the impacts of culture on the frequency, occurrence, and types of hand gestures in multimodal interfaces.

Bischel et al. [3] investigated a designer describing a mechanical device to another designer. Similar to their experiments, we videotaped and analyzed the gestures of the participants from two different cultures describing two different chairs using their speech and hands. We segmented the video records using timestamps, and analyzed them using metrics such as frequency, occurrence, and the quantity of certain gesture types.

The paper is structured as follows: Section II provides a review of related literature on gesture and culture. Section III presents the experiment conducted. Section IV analyzes the data collected, and the results of frequency, occurrence and type of gesture amongst the samples. Finally, Section V discusses the experiment results and relation with culture studies, and Section VI presents our conclusions drawn from the results.

II. LITERATURE REVIEW

The aim of Human Computer Interaction (HCI) is making interactions as natural as possible, as if communicating with another human [4]. Gestures are used as a way of expression in interaction, either with or without a human. There is evidence that gestures convey information *redundant* to the information conveyed in speech [5] and gesture is a precursor to speech. Humans have an innate need to use gestures; since they complement our ideas, to such an extent that humans are known to gesture even when talking on the phone [6].

The means to communicate with computers have evolved from classic mouse input, to rich multimodal data [7]. Multimodal interfaces, that use multimodal data, have combined various user input modes beyond the known keyboard and mouse input/output [8]. They now include a wide range of interaction methods; such as hand gestures (both static and dynamic), as well as speech, and head and eye tracking. Most recently, Razer™, the world leader in high-performance gaming hardware, launched an adaptive tactile keyboard with a switchblade user interface, suggesting only the recurrent need to provide players with more options. Games and interactive entertainment industries are not, however, the only application areas for gesture based interfaces. More serious applications exist, such as The Intuitive da Vinci surgical system that is used for the capture of subtle motions of a surgeon, to teach novices complex procedures [9]. One may assume that in tasks such as the manipulation of objects, cultural implications might not be of considerable importance, but in the context of cultural and physical differences between surgeons, the topic calls for more attention [10].

Gesture-based interfaces enable freer, more intuitive, and richer digital interactions than conventional user interfaces [11]. These enhanced interactions lead to better idea generation [12]. With the decrease in the price of sensors and the growth of processor capacity, an interesting step in multimodal interface design is the creation of natural and invisible interfaces that are called ambient gestures [13]. These invisible interfaces are designed to support ubiquitous interactions with everyday computing technology. Such examples include lowering the volume of a stereo from afar using a hand gesture. This way the user does not have to leave their original activity to perform an action adding the value of ubiquity. Through appropriate gestures, these technologies allow immersion, navigation and interaction to support idea generation [14].

A. Gesture classification and segmentation

Many gesture recognition systems have been technology driven and address gesture tracking needs, rather than the requirements of human behavior [15].

When developing gesture based interfaces, programmers and designers work together to understand what types of gestures are most frequently used for interaction. Therefore, there have been many attempts to design appropriate gesture classification and segmentation "dictionaries". However, none of them provide a complete guide for what gestures are mostly used and by what group of users.

Gestures offer versatility when representing objects, or qualities of these in the scientific domain. The main problem here is that there is no common database of gestures that can be used by both developers and designers. At the top level, they are divided into representational and non-representational gestures. Representational gestures represent physical things being conceptualized by the person gesturing in an abstract or physical manner. Gesture also allows representation of movement through imitation [6]. The non-representational gestures, accompany or stress *speech*, and also involve space relations, like pointing. In a way, they are culture-specific emblems, but *how* culture-specific are gestures remains as a question to explore.

Gestures have also been classified according to their purpose. They could be goal oriented (change of position, shape), empty handed (wave, snap, point, take), for indirect manipulation (set, stop) or haptic exploration (touch, stoke, knock) [14]. The most recognized gesture classification is the one established by McNeill in 1992 [16]. McNeill classifies 4 types of gestures (See Table 1): iconic (resemble what is being talked about, e.g., flapping arms when mentioning a bird), metaphoric (abstractedly pictorial, e.g., drawing a box shape when referring to a room), beat (gestures that index a word of phrase e.g., rhythmic arm movement used to add emphasis), and deictic (gestures pointing to something, e.g., while giving directions).

The iconic gestures are of particular interest to HCI as they allow accurate depiction of objects encountered by the user. An important issue here is the presence of transition movements, or junk gestures, as these are classified as meaningless, since they do not convey information.

When training a model for gesture recognition, it must be done by segmenting the individual gestures and then interpreting their representation individually or as a whole, depending on the model. A common motion segmentation technique is using distance signals to determine the type of gesture on the basis of the contour of the gesturing person's body or hand [17]. Li and Greenspan in [18] focus on how the endpoints are located. In order to do this, they had participants repeat various actions several times in order to document the variances. These variances, they claim, are useful for identifying the range of a given gesture, and therefore, provide a better identification.

The most common errors in gesture segmentation are mainly classified as three types [18]: substitution errors, deletion errors and insertion errors. Substitution errors appear when an incorrect gesture is substituted for the correct one. Deletion is where a correct gesture is omitted in the recognized sequence. Insertion is when an extra gesture was added in the sequence; hence, instead of having one gesture. Here two gestures are segmented (Gesture A and B) instead of one, because the technique recognizes part of A as a separate individual gesture, B.

Relatively less explored issues in gesture segmentation are:

- how to detect the differences between a dynamic gesture (where the path of each limb is relevant), and a static posture (where only one particular position of the limbs is relevant),
- distinction of feature descriptions (which features optimally distinguish the variety of gestures and postures from each other and make recognition of similar gestures and postures simpler), and
- gesture meaning identification (what do certain gestures mean, how can they be reliably interpreted so that the correct actions are undertaken).

B. Culture

Hooler and Beattie [19] claimed that speakers draw on gestures in order to fulfill particular communication functions. Sometimes, culture is a boundary for interpersonal communication; in the same manner

TABLE 1. GESTURE CLASSIFICATION DETAIL

Gesture	Atributes	
	Function	**Linguistic example**
Iconic	Resembles that which is being talked about	Flapping arms like wings when talking about a bird
Metaphoric	Abstractly pictorial; loosely suggests that which is being talked about	Making a box shape with hands when talking about a room.
Beat	Gestures with only two phases (up/down, in/out) indexing the word or phrase it accompanies as being significant	Rhythmic arm movement used to add emphasis
Deictic	Gestures pointing to something or somebody either in concrete or abstract	Pointing while giving directions

Source: Michael Berry, The importance of Bodily Gesture [15]

technology may enable or even hinder a communication style inherent to a culture [20]. Metrically, culture could be reflected in the interactivity, symbol variety, rehearsability and pre-processability of gestures. Therefore, the cultural background might be an influential factor in the design of gesture-based interfaces.

Technology has been conceived in 'prosthetic' terms, as an extension to the body, or support for tasks [21] and given the global diversity; cultures will perceive these tasks differently. Language and representation are critical elements in the study of culture, because humans are locked into their cultural perspectives and mindsets [22]. As defined by Hofstede [23] "Culture is the collective programming of the mind that distinguishes the members of one group or category of people from another".

1) Culture and Interfaces

Humans communicate and exchange information with a system through interfaces. The more familiar or intuitive an interface is, the higher its usability. Maximum usability of a system can be obtained through the appropriate design of support-focused interfaces.

Cultural preferences determine the type of layout, texture, pattern and color [24] in website portals. Certain colors are offensive or uncomfortable for certain cultures. For instance, red is bad luck for Koreans, therefore, Korean websites might avoid the use of red. Given the example, it is noted that to attract a targeted market, there is a need to adapt interfaces to a specific culture. Culture does not exist as a computational term in HCI, even though, as revealed, there are efforts to tailor interfaces. With any use of the technology, the success depends on the capabilities embedded in a persona who is "programmed" in a specific way. The mental "coding" of this persona will affect the usability of the system as well as its interface.

The cultural behavior is generally perceived visually, but it is not always evident until there is an interaction. Rehm, Bee, and André [25] tried to identify the culture of the user so that the behavior of an interactive system could be adapted to culture-dependent patterns of interaction. This was achieved via a Bayesian network model based on gesture expressivity and made use of metrics, such as speed, power or spatial parameters.

A study comparing North Americans to Chinese [26] explored gesture frequency of bilinguals when speaking in both languages. The study demonstrated that American monolinguals used more gestures than Chinese monolinguals and suggested that the American culture is a high-gesturing one. The study also noted that Chinese bilinguals used more gestures than Chinese monolinguals, and suggested that there was possibly a transfer in gesture frequency. The study stated that language is a medium by which culture is transmitted. When speaking in a particular language, the speaker might display the gesture pattern found in the corresponding culture. This means that bilinguals from a lower-gesturing culture and language (such as Chinese) used more gestures when talking in the secondary high-gesturing language (English). In this particular scenario, the difference in gesture frequency might be attributed to the culturally varying attitudes towards body movement.

Some research studies focused on the variation in the expressions between cultures as well as the influence of the language. Nicoladis explained that, this could be due to the fact that bilinguals might tend to produce more gestures, when speaking in a language they might feel weaker at [26]. In this case, the ability to express themselves via gestures might help the speaker to break down the preverbal, spatial-motoric information for verbalization.

In remote international collaboration, gesturing is also a concern. Here, each participant has their own symbolic, iconic and metaphoric influence on their gestures [28]. Given that the Internet is a technological, cultural, political and economic phenomenon, it produces an extraordinary volume of cultural expression [21]. There is a merge of culture amongst the latest generations because computers, interfaces and common systems increasingly assist our lives and have an influence on our gestures. There are studies stating that, since the latest generations have grown up with a mouse based interface, this has a more standard and global effect on command gestures used in their interactions [29]. This suggests the influence of interfaces on culture as opposed to the goal of having culture to influence interfaces.

As stated by Hofstede regarding communication technologies, the software of the machines may be globalized, but the software of the minds that use them is not [23]. Therefore, the dominance of technology over culture is an illusion, and differences between cultures exist.

2) Hofstede's Cultural Dimensions

Hofstede [29] has developed a set of cultural parameters that describe the way in which national societies are built and the rules by which people think, feel and act. These differences are defined as five dimensions and are measured as indexes. The higher or lower the index, more or less the culture portrays this feature.

The Hofstede model of dimensions of national culture has been applied predominantly in international business; marketing and consumer behavior works [30]. The brief descriptions of Hofstede's dimensions are as follows:

Power Distance (PDI) is the acceptance and expectation of power to be distributed unequally viewed from the less powerful members of organizations and institutions (like the family). The higher the index, the more a society views an unequal distribution of power as relatively acceptable (as in Malaysia, Philippines, and Mexico). A lower index indicates that the unequal distribution of power is relatively unacceptable (Austria, Denmark, NZ, Ireland).

Uncertainty Avoidance (UAI) indicates the extent to which a member of society feels uncomfortable or comfortable in ambiguous or abnormal situations. A low UAI index indicates that the society is comfortable in unpredictable situations and has a high tolerance for ambiguity (as in Denmark, Singapore, China, and Sweden). A higher index here means that people prefer predictable situations and have a lower tolerance for ambiguity, a factor that is normally reflected in the abundance of laws in these countries (as in Belgium, Salvador, Greece, and Guatemala).

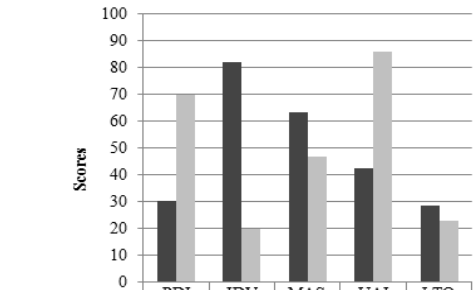
Individualism (IDV) is the extent to which individuals are merged into groups. A high IDV index means members in the society define themselves as individuals and form looser ties with their groups and immediate families (in USA, Anglo-Celtic, UK, and Canada). Countries with lower IDV have stronger bonds to their groups and extended families. Here a group membership forms a person's self identify (in Guatemala, Ecuador, Indonesia, and China).

Masculinity (MAS) refers to the distribution of emotional roles between the genders, and also serves to classify a culture as assertive and competitive (masculine) or modest and caring (feminine). A country with a high MAS index, values achievement and competitiveness, as well as acquisition of money and other material objects (as in Slovakia, Japan, Austria, and Venezuela). In low MAS cultures people value the maintenance of good relationships, are more modest and people oriented, caring for the quality of life (like in Norway, Netherlands, Sweden, and Chile).

Long-Term Orientation (LTO) considers that countries high on this index foster pragmatic virtues oriented towards future rewards, in particular saving money, persistence, and adapting to changing circumstances (as in China and India). Countries with lower LTO are short-termed and give more importance to attitudes towards national pride, tradition, and fulfilling social obligations (as in United States and Norway).

The cultures used in the experiments for this paper are Anglo-Celtic (Australian, British, Irish, New Zealanders) and Latin Americans (American countries where Spanish is primarily spoken: Argentina, Chile, Colombia, Costa Rica, Ecuador, Salvador, Guatemala, Mexico, Panama, Peru, Uruguay, and Venezuela). Fig. 1 shows a comparison of the cultural dimension indexes from both samples; an average was taken of the indexes of the countries mentioned above. The Anglo-Celtic culture had a lower PDI (30 to 70), and UA (43 to 86). On the other hand, it had a higher IDV (82 to 20), MAS (63 to 47) and LTO (29 to 23) than the Latin American one.

Therefore, one can assume that due to the greater equality (Low PDI) characteristic Anglo-Celtics have, they are more individualistic (High IDV) and can master new challenges (Low UAI) better than their fellow Latin American colleagues. Hosftede developed a solid foundation for identifying the possible complication of



	PDI	IDV	MAS	UAI	LTO
Anglo-celtic	30	82	63	43	29
Latin American	70	20	47	86	23

Figure 1. Hostedes 5D Model comparing Anglo-Celtic and Latin American countries.

cross-cultural interactions, as well as what creates cultural differences and how they would act upon this [20].

Even though Hosftede is cited by an extensive amount of sociologists and anthropologists, for the analysis in this study, it is also beneficial to analyze the context classification made by the anthropologist Edward Hall [22]. Hall identifies a culture's use context in routine communication and classifies them as High or Low. In a high context culture (including much of the Middle East, Asia, Africa, and South America), many things are left unsaid, letting the culture explain. There is more non-verbal communication, a higher use of metaphors, and more reading between the lines. In a lower context culture (including North America and much of Western Europe), the emphasis is on the spoken or written word. They have explicit messages, focus on verbal communication, and their reactions are visible, external and outward [31].

Anglo-Celtic cultures (e.g., Australian, British, Irish, and New Zealanders) categorize as low context cultures and Latin Americans (American countries where Spanish and Portuguese are primarily spoken) correspond to the high context cultures. This classification lets us make certain assumptions, like the Anglo-Celtic may predominantly use words, while the Latin Americans would use gestures.

The characteristics identified for each of the samples (IDV, UAI, MAS) will be later referred to in order to interpret their influences on gesture behavior after the experimentation.

III. EXPERIMENT

A set of experiments, following Bischel's experiment guidelines, were conducted in order to explore the influence of culture in gesture behavior. The participants were required to describe two chairs before a camera (See Fig. 2). Bischel's experiments bring language and gesture together; both of these are important in defining a culture. From this moment on, the classical chair will be referred to Chair 1 and the abstract chair will be referred to as Chair 2. Throughout this study, the observational task analysis method was used. Video analysis technique permits a

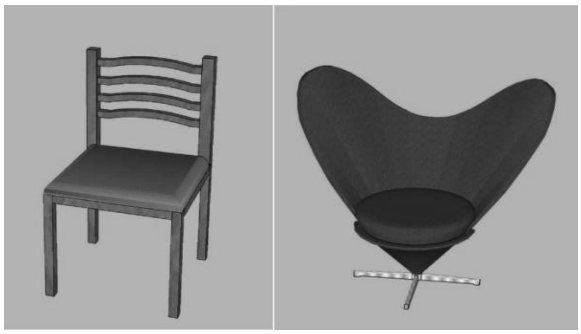


Figure 2. Classic chair (left) CHAIR1 and Abstract chair (right) CHAIR2

careful analysis of gestures occurring at certain timestamps. This is helpful in identifying individual differences in gesture behavior.

A. Hypothesis

The paper tries to prove that: "Designers' culture may affect gesture recognition in multimodal interfaces because of variations in gesture type, gesture frequency, and gesture occurrence".

This hypothesis brings together the subjects of Section II (gesture, multimodal interfaces, gesture segmentation and culture based theories). The three metrics stated in the hypothesis are gesture type, frequency and occurrence.

- Gesture Type. The gesture type is based on McNeill's classification. Certain types of gestures could be attributed to different cultures; therefore, it is important to analyze the type of gesture that is mostly performed.
- Frequency. The frequency is measured as the number of gestures performed by a participant divided by the period of the gesture of the same participant. This way one can obtain the gestures per second, which will help assess speed of gesture performance and point out what gestures are most significant for a gesture recognition system.
- Occurrence. Occurrence measures the appearance of the gestures. This helps to identify if certain gestures are culture-oriented or task-oriented (i.e., related to the task being performed).

B. Experiment Guidelines

Participants were encouraged to use as many gestures as possible, just as in Lui and Kavakli "Temporal Relation between speech and co-verbal iconic gestures in multimodal interface design" [32]. The analysis methodology is via video analysis using a video annotation tool called Anvil (Fig. 3).

1) Procedure

In order to obtain a detailed description, and for us to derive data from video recording, the participants were required to sit in front of a camera in an enclosed setting. This framework helped avoid distractions and background noise that will help with the proper gesture classification.

The influence on gesture performance when holding a seating or standing position by the gesturer has not been assessed. The seating position was thought to simulate seating in front of a screen, where the actual interaction could take place if the user was using a virtual reality interface.

They were instructed that this was a cultural experiment that required analyzing gestures they used to describe a given object. They were encouraged to use both hands, to use as many gestures as possible, and take the time they considered adequate.

2) Participants

There were a total of 8 Latin American participants and 11 Anglo-Celtics videotaped, but only the ones with clearer hand gestures and comprehension of the task were chosen. A criterion for deselecting a video footage for analysis was either the lack of gestures, or the lack of iconic gestures, which are the focus of this study.

The final selection was 5 participants from each sample group, totaling 10 participants. For the purpose of the experiment, two samples were needed, one with English as a first language (Anglo-Celtics), and one with English as a second language (Latin Americans). For the second sample, it was important that they were sufficiently proficient and immersed in an English speaking country (Australia) at least for the past 6 months.

3) Gesture coding

For gesture analysis, the video footages were analyzed and then segmented (Fig. 3). For each occurrence, the gesture type (repetition, beat, iconic, metaphoric, deictic, and junk) that was performed by the participant was recorded. These correspond to McNeill's classification, but the repetition gesture (which is a type of deictic gesture) was coded separately because of the assumption made by Nicoladis in [27] that states that bilinguals might produce more gestures when speaking in a language they might feel weaker at.

Therefore, repetition was considered to be a potential factor that reflects culture, as uncertainty in the language, or description, could be channeled this way. Junk gestures

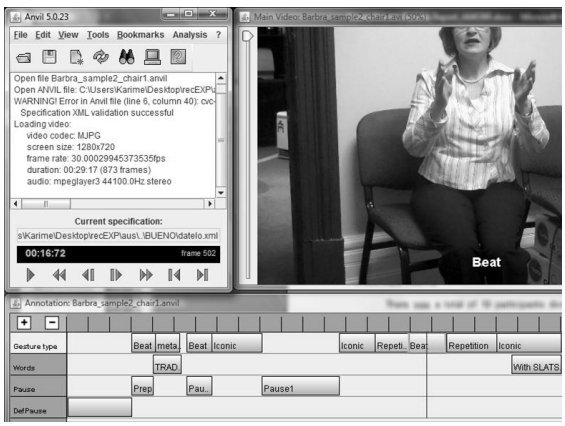


Figure 3. Anvil Snapshot of Anglo-Celtic observation

were identified as gestures without a particular meaning. This could be a gesture that the user takes the gesture back, (which is a "mistake") or made some transition movements.

Gestures are separated by pauses, and a pause is defined as a temporary stop in action or speech [33]. The purpose of this pause was to eliminate the period of inactivity at the beginning of a video, when the participant explains what he or she might do, or when the participant states that he or she has ended the gesture.

4) Speech Coding

As a result of verbal descriptions (words) used to give meaning to the participants' depiction, distinctive words were coded accompanying the gestures. These words, which were also coded on Anvil, were hand written on the track "words". The words coded on this track were identified as significant because they were especially stressed. The classification of words was as follows: adjectives, parts of the chair, verbs, positions and shapes.

The rationale behind this structure is to identify, which accompanying words were used to express gestures so that comparisons could be drawn between the two cultures. The condition where a specific gesture is expressed with a word or not, was not investigated, as the focus was limited to type, frequency, and occurrence of gestures.

IV. RESULTS

Approximately 10 minutes of monologue object descriptions in video footage was obtained. Table 2 contains the metrics referred to in this section. Seconds were used as the time measuring unit. In total, there were 595.52 seconds of video footage captured in 17754 frames. The video footage was composed of 20 individual videos, 2 videos per each participant and a total of 231 gestures.

A. Chair 1.

Chair 1 was a traditional chair, with common characteristics that all participants were able to relate to. When analyzing the data, it was found that Anglo-Celtics on average scored higher values in the gesture duration (1.84s), number of gestures (65), and gesture time per person (22.74s). It is also noticed that the Anglo-Celtics had a higher share (52%) of gestures recorded. The sample shows a considerable variation from one participant's representation to another's, as the mean was 12.8 gestures and the standard deviation (SD) was 5.63.

On the other hand, Latin Americans on average scored lower the values in the gesture duration (1.49s), number of gestures (59), and gesture time per person (17.81s). The mean was 1 gesture lower (11.8), but the variation in this sample was less (2.16) than that of the Anglo-Celtics, meaning more uniformity amongst the sample.

Chair 1 results reveal that Anglo-Celtics produced higher number of gestures in a longer period. Latin Americans, on the other hand, produced fewer gestures that were faster and shorter. The gesture type that was most recurrent was iconic, followed by repetition, beat, metaphoric, junk and then deictic. Anglo-Celtics had more repetition, iconic, and metaphoric gestures. Latin Americans maintained higher deictic and junk gestures. In matters of occurrence, only one Anglo-Celtic performed a deictic gesture.

Frequency might look to be proportional with the amount of gestures but in fact it was not. The overall frequency was higher with the Latin Americans, even though the amount of gestures was less when compared to the Anglo-Celtics (65 to 59). Taking the iconic gestures, for example, Anglo-Celtics on average produced more (6.2) gestures, but their frequency was lower (0.26), as they performed these in a longer time.

In regards to verbal coding, the words used to describe Chair 1 were 28 in total (See Table 3). Most of them correspond to adjectives (See Table 4) given to the chair as a whole, as well as specific parts of it. From this 9 words in total were uniquely used by Anglo-Celtics, and 13 by Latin Americans. Both shared 6 specific words: horizontal, front, seat, legs, back and square. Therefore, it can be concluded that the chair was easy to describe both physically and qualitatively.

B. Chair 2.

Chair 2 was a more complex chair, with a more elaborate structure. To an extent it is considered as "abstract". All participants mentioned some way the chair was different, therefore more difficult to describe. The data for the gesture representation for this chair is also in Table 2. The mean of gestures for the chair 2 descriptions for both samples was 10.8 (Total number of gestures per chair divided by the 10 participants) and the SD was 5.65

TABLE 2. VIDEO ANALYSIS FOR CHAIR 1 AND CHAIR 2

Chair	**Metrics**								
	Total no. gestures per chair	***Sample***	***Average gesture duration***	***Total no. of gestures per chair per sample***	***Percentage attributable to sample***	***Average gestures per person in sample***	***Standard deviation (SD)***	***Average gesture time per person in sample***	***Frequency of gesture performance per sample***
Chair 1	124	Anglo-Celtic	1.84	65	52%	12.8	5.63	22.74	0.56
Chair 1		Latin American	1.49	59	48%	11.8	2.16	17.81	0.66
Chair 2	108	Anglo-Celtic	1.73	65	60%	13	7.17	23	0.56
Chair 2		Latin American	1.67	43	40%	8.6	2.88	14.22	0.60

(Average of both SD's). This SD is higher than Chair 1, showing different approaches followed by participants to describe the chair.

The data reveal that Anglo-Celtics on average scored again higher values in gesture duration (1.73s), number of gestures (65), and gesture time per person (23s). Anglo-Celtics had higher scores in the average duration of gestures and had a higher share (60%) of gestures recorded. Latin Americans on average scored lower values in gesture duration and number of gestures, but these values were significantly lower than the time required for Chair 1.

As with gesture and frequency, the differences in chair descriptions are also reflected in the use of words. Chair 2 surpasses Chair 1 with a total of 30 words (See Table 3). Anglo-Celtic uniquely used 13 words and Latin Americans 10. Both samples used 5 words in common: Back, seat, curved, heart, and cone. This variation could be because the chair was more complex and needed a higher evaluation and explanation. It was found (Table 5) that when describing the back of the chair, the same word was used, and this word was "heart". It was also found that there were more verbs and shapes in this chair description; words used to describe the chair metaphorically and qualitatively.

Chair 2's general data confirm that Anglo-Celtics again had higher number of gestures in a longer period. This is reflected in the lower frequency achieved by the Anglo-Celtics in Fig. 4b. Latin Americans on the other hand performed fewer gestures that were faster and shorter. Table 2 shows that the Anglo-Celtics used more or less the same amount of gestures as before. This could be because they are more comfortable when describing the "abstract" concepts in their first language.

TABLE 3. WORDS DERIVED FROM ALL CHAIR RECORDINGS

Sample	Recorded Words			
	Anglo-Celtic	***Latin American***	***Both***	***Total***
Chair 1	9	13	6	28
Chair 2	13	10	5	30

TABLE 4. WORDS DERIVED FROM ALL CHAIR 1 RECORDINGS

Classification	Recorded Words
Adjectives	Traditional, Thin, Arched, Vertical, Squarish, Long, Curved, Rectangular, Straight, Round, Normal, Flat
Parts	Slats, Bottom, Front, Seat, Legs, Place, back
Verbs	Support
Position	First, left
Shapes	Holes, squares, sticks, stripes

TABLE 5. WORDS DERIVED FROM ALL CHAIR 2 RECORDINGS

Classification	Recorded Words
Adjectives	Round, Horizontal, Perpendicular, Funny, Curvy, retro, stainless, steal, shortest, curved, symmetric
Parts	Back, Bottom, Down, seat
Verbs	Focus, Spread, Crosses, extend
Position	Middle
Shapes	Heart, Cone, Square, Circle, Wing, Stick, Oval, Peak, Cross, triangle

C. Findings

After analyzing the performance of both samples, in this section, the results of the metrics stated in the hypothesis (gesture type, frequency and occurrence) are delivered. The results come as follows:

1) Frequency

Gesture frequency indicates that overall the Latin American sample performed more gestures per second; however, this evidence is not enough to state which culture was more expressive than the other. The use of gestures involves various factors, such as the comfort of a person had in front of the camera, or the confidence with the object being described, as well as the language. Chair 1 had iconic and repetition gestures with higher frequency in both samples, and this is reflected in Fig. 4.a and Fig. 4.b. Chair 2 on the other hand had an increase in junk and metaphoric gestures. The most significant gestures for gesture recognition are the iconic ones as well as repetitions, and subsequently they are the ones that convey the description of the chair more significantly.

2) Occurrence

There are no junk and deictic gestures in the description

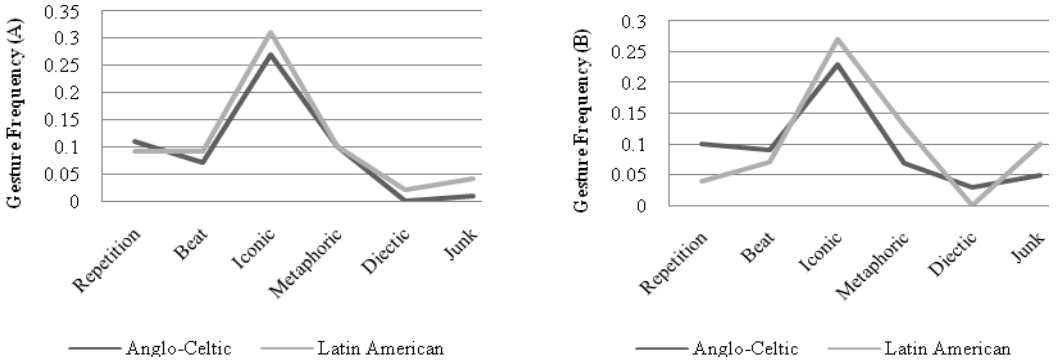


Figure 4. Gesture Frequency of gestures by chair and sample:
Gesture Frequency Chair 1 (a)
Gesture Frequency Chair 2 (b)

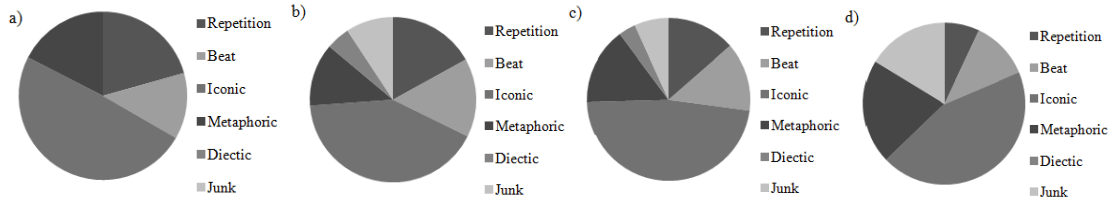


Figure 5. Gesture types by chair and sample:
Chair 1- Anglo-Celtic (a) Chair 1- Latin American (b)
Chair 2- Anglo-Celtic (c) Chair 2- Latin American (d)

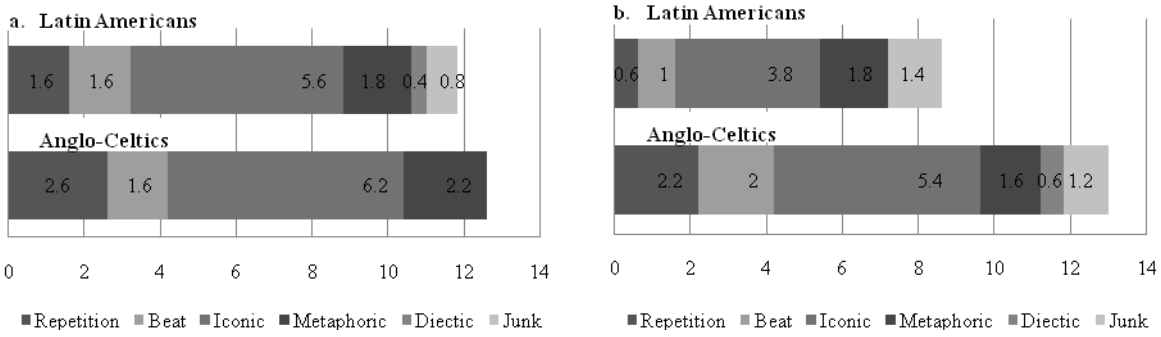


Figure 6. Average Gesture type by sample in
Chair 1 (a) and Chair 2 (b)

of Chair 1 for the Anglo-Celtic sample, but they do appear in Chair 2 when the number of gestures increases. This means that the occurrence of gestures was related to the task, not to the culture. This may be because Chair 2 was more complex and there was a need for more explanation by the user. This explanation required more cognition and the extra gestures reflected this process.

3) Gesture Type

For Chair 1, the iconic gestures were close to 50% in both sample groups (See Fig. 5a and Fig. 5b), showing equality. For Chair 2, the iconic gestures diminish and metaphoric gestures increase for the Latin American sample group (See Fig. 5c and Fig. 5d). Again, this may be related to the relative complexity of the chairs. Regarding verbal depictions, each iconic and metaphoric gesture was related to at least one word, reflecting the participants' cognition.

D. Comparisons

In this section, we summarize the important points described in Section C. Numerically; Anglo-Celtics did not display too much variation between chair descriptions, even though they all encountered a new and different chair that required a lot more extraordinary description.

Chair 2 had more gestures on average by participant in the Anglo-Celtic sample. On the contrary, in Chair 2, Latin Americans performed fewer gestures on average by participant. The reason could be the degree of comfort Anglo-Celtics had when describing an abstract chair.

The SD was again higher with the Anglo-Celtics. This made it hard to identify a pattern. On the other hand, Latin Americans had a smaller SD and more frequent gestures, meaning shorter, more concise and common gestures by most of the participants.

The gesture frequency was higher in Chair 1, and it increased with the Latin Americans. This could be partly because; they scored higher values in junk gestures in the description of Chair 2. Latin Americans had more frequent gestures in both chairs. This means that they performed more gestures per second, even though they had fewer gestures in total. The smaller count of gestures by Latin Americans is justified by shorter time frame in which they performed the gestures.

Deictic and junk gesture occurrence was significantly less than the other gestures. Therefore, the idea of Latin American's being more explanatory with hands is disproved. A potential explanation in this case, is perhaps the implication of having to speak in a second language.

Given the distribution of gestures (Fig. 5 and Fig. 6), it is identified, that in general, iconic gestures decrease with Chair 2, as well as the repetition gestures. In contrast, junk and deictic gestures appear more; the average of these gestures therefore, increases from one chair to another (Fig. 6a and Fig. 6b).

Latin Americans used more distinctive words for Chair 1 (13) and less in the Chair 2 (10) (Table 3). Using less gestures and words to describe Chair 2 could probably mean a better selection of words and gestures, or the lack of vocabulary. The higher word count for Chair 1 may refer to either a higher degree of confidence, or more predictable and well structured ideas.

V. DISCUSSION

Now the relation between the gesture metrics and the cultural attributions made by both Hofstede and Hall (Section II b) is presented. As Anglo-Celtics are low context cultures, they used more words and gestures in longer time, since they took time to explain the chair in detail. On the other hand, Latin Americans, which represent the high context culture, performed fewer gestures, in shorter time and used fewer words. The element that calls for attention is the higher use of metaphoric gestures, a trait of a high context cultures. This exemplifies a characteristic of a society that relies on reading between the lines and letting nonverbal cues explain the meaning.

The cultural analysis now continues by relating gesture performance with Hofstede's cultural dimensions. A clear integration between the culture literature cited and the experiment conducted is exposed in Table 6. As mentioned before, the traits that are mostly reflected are IDV, UAI, and MAS.

- IDV. This trait could be related to the fact that the SD between samples is higher with the Anglo-Celtic cultures reflecting the societies high individualism index (IDV, 82). On the other hand, the low SD with the Latin Americans shows the low individualism index (IDV, 20).
- UAI. This trait could be reflected in the overall impression of Chair 2. The Anglo-Celtic sample did not vary too much in gesture means and time from one chair to another, showing greater comfort with adverse situations (UAI, 43). It is possible to say that Latin Americans showed their high uncertainty avoidance (UAI, 86) since they use less time and limited gestures, possibly sticking to "what they knew" instead of managing the abstract.
- MAS. This trait could be related to the fact that the Anglo-Celtics as a low context culture are more masculine and assertive (MAS, 63), in comparison to the Latin Americans that are more human-oriented and feminine. Therefore there is a higher use of metaphors (MAS, 47) in their descriptions.

It is important to remark that the Latin Americans in this sample have more of an advantage with the language compared to "at home" Latin Americans, as they have been immersed in a different culture and language for the past 6 months. Regardless of that, they still performed fewer gestures and chose different words.

VI. CONCLUSION AND FUTURE WORK

The goal of this paper was to explore if gesture-based interfaces could be affected by a user's culture. The literature review agrees that any interaction is a result of user, task and input. Aside from performance or stability issues, gesture-based interfaces are subject to a context problem. In the international scene, depending on where participants are from, their style of communication will vary. This analysis arrived to the conclusion that as the complexity of a task increases, so does the use and type of gestures. The metrics stated in the hypothesis influence multimodal interfaces and their performance in the following ways:

- Frequency may affect the recognition rate because of the need for faster and more efficient algorithms.
- Occurrence also affects interaction due to the possibility of absence (zero occurrences) of certain gestures that may convey functionality (i.e., iconic).
- Gesture type, as well as occurrence, also affects the goal that the user wishes to attain. Identifying and classifying certain gestures due to their use during trials would permit the identification of type tendencies and will assist in embedding differences in the development of the gesture recognition tools.

Due to the "freedom" that hand gestures provide, gesture based interfaces gain popularity. The aim of HCI is to have users adopt the new technologies for interaction because their usability is better. Studies have shown that culture influences a user's acceptance of the technology. A more conservative or traditional culture, such as the Latin American culture, could take more time to adapt. This was visible with the differences in frequency rates between the classic and abstract chairs. The performance of a gesture based interface will not only be affected by the task being performed, but by the cultural background and language skills of the user. Therefore, the design of gesture-based interfaces not only requires a multidisciplinary approach, but also a culturally sensitive one. The work conducted in order to develop applications that are consistent with the perception, use, and understanding of gestures by users, still continues.

It is acknowledged that future studies need a larger sample size. Similarly, future studies need to focus on the consistency of the annotations by having more than one coder to increase the objectivity of analysis. The results might have significant variations if the experiment were to

TABLE 6. INTEGRATION OF EXPERIMENT AND CULTURE

Sample	**Metrics**			
	Context	***Predominant culture trait by Hofstede***	***Metric Evidence***	***Predominant Gesture Type***
Anglo-Celtic	Low context (assertive, rely con words)	Individualism Masculinity	High SD Constant gestures beween chairs More gestures and more time	Iconic
Latin American	High context (rely heavily on non verbal communication)	High Uncertainty Avoidance Collectivism	Low SD Fewer Gestures in the second Chair Fewer gestures in less time	Metaphoric Repetition

be carried out in Spanish, the native language of the Latin American culture. Further research studies may also attempt to investigate the effects of gender on gesture behavior.

ACKNOWLEDGMENT

This research is supported by the Australian Research Council Discovery grant DP0988088 to Kavakli, titled "A Gesture-Based Interface for Designing in Virtual Reality". Also, the experiment on which this project was based, is inspired by Jing Liu's previous work titled "Temporal Relation between speech and co-verbal iconic gestures in multimodal interface design". Authors are grateful to her for her support in sharing data.

REFERENCES

[1] Kavakli, M. and Nasser, K. 2012. Influence of Culture in Multimodal Interfaces: Differences in Gesture behavior between Anglo-Celtics and Latin Americans, *MMEDIA 2012: the fourth International Conferences on Advances in Multimedia*. Pp 148-154 ISBN: 978-1-61208-195-3

[2] CARLA. 2012. What is Culture?. The Center for Advanced Research on Language Acquisition, http://www.carla.umn.edu/culture/definitions.html [Accessed: December 15, 2012]

[3] Visser, W. and Maher, M.L. 2011. The Role of gesture in designing. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 25, pp. 213-220. Cambridge University Press, doi: 10.1017/S08900604/11

[4] Wachs, P., Kölsch, M., Stern H., and Edan Y. 2011. Vision-Based Hand-Gesture Applications. *Communications of the ACM*, 54 (2), pp. 60-71, doi:10.1145/1897816.1897838

[5] Kita, S. (ed.). 2003. Pointing and Placing. Pointing: Where Language, Culture and Cognition Meet, http://www-psych.stanford.edu/~herb/2000s/Clark.Pointing.placing.03.pdf [Accessed: December 15, 2012]

[6] Bavelas, J., Gerwing, J., Sutton, Ch., and Prevost, D. 2008. Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, 58 (2), pp. 495-520 ISSN: 0749596X, doi: 10.1016/j.jml.2007.02.004

[7] Gullberg, M. 2010. Methodological reflections on gesture analysis in second language acquisition and bilingualism research. *Second Language Research*, 26 (1), pp. 75–102. doi: 10.1177/0267658309337639

[8] Karray, F., Alemzadeh, M., Saleh, J., and Arab, N. (2008) Human-Computer Interaction: State of the Art. *International journal on smart sensing and intelligent systems*, 1(1), http://www.s2is.org/Issues/v1/n1/papers/paper9.pdf [Accessed: December 15, 2012]

[9] Lanfranco, A., Castellanos, A., Desai, J., and Meyers, W. 2004. Robotic Surgery. *A Current Perspective Ann Surg*, 239 (1), pp. 14–21. doi: 10.1097/01.sla.0000103020.19595.7d

[10] Purdue University. 2011. Future surgeons may use robotic nurse, 'gesture recognition'. *ScienceDaily*, http://www.sciencedaily.com/releases/2011/02/110203152548.htm [Accessed: December 15, 2012]

[11] Van den Hoven, E. and Mazealek, A. 2010. Grasping Gestures: Gesturing with physical artifacts. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 25, pp. 255-271, doi:10.1017/S0890060411000072

[12] Kim, M.J. and Maher, M.L. 2008. The impact of tangible user interfaces on spatial cognition during collaborative design. *Design Studies*, 29 (3), pp. 222–253, doi: 10.1016/j.bbr.2011.03.031

[13] Karam, M., Hare, J., and Schraefel, M. n.d. Ambient Gestures, http://users.ecs.soton.ac.uk/amrk03r/AmbientGestures.pdf [Accessed: December 15, 2012]

[14] Kavakli, M., Taylor, M., and Trapeznikov, A. 2007. Designing in virtual reality (DesIRe): a gesture-based interface. In *Proceedings of the 2nd international conference on Digital interactive media in entertainment and arts* (DIMEA '07). ACM, New York, NY, USA, 131-136, doi:10.1145/1306813.1306842

[15] Kavakli, M. 2008. Gesture Recognition in Virtual Reality, *Special Issue on: "Immersive Virtual, Mixed, or Augmented Reality Art" of The International Journal of Arts and Technology* (IJART), 1 (2), pp. 215-229. doi: 10.1504/IJART.2008.021928.

[16] Berry, M. 2009. The Importance of Bodily Gesture Example 1. McNeill's classification of gesture types, http://www.mtosmt.org/issues/mto.09.15.5/berry_ex1.html [Accessed: December 15, 2012]

[17] Kang, H., Lee, C., and Jung, K. 2004. Recognition-based gesture spotting in video games. *Pattern Recognition Letters*, 25 (15), pp. 1701-1714, doi:10.1016/j.patrec.2004.06.016

[18] Li, H. ad Greenspan, M. 2011. Model-based segmentation and recognition of dynamic gestures in continous video streams. *Pattern Recognition*, 44, pp. 1614-1628, doi:10.1016/j.patcog.2012.12.014

[19] Holler, J. and Beatiie G. 2004. The interaction of iconic gesture and speech in talk. *Gesture Based Communication in Human Computer Interaction*, 2915, pp. 3853-3885, doi: 10.1007/978-3-540-24598-8_6

[20] Massey, A., Hung, C., Ramesh V., and Montoya-weiss, M. 2001. Cultural perceptions of task-technology. *Communications of the ACM*, 44 (12), doi:10.1145/501347.501353

[21] Murphie, A. and Potts, J. 2003. Culture and Technology. N.Y: Palgrave Mcmillan

[22] Hall, S. 2007. Review: Representation: Cultural Representations and Signifying Practices (Culture, Media and Identities Series) London: SAGE Publications Limited

[23] Hofstede, G. H., Hofstede G. J., and Minkov, M. 2010. Cultures and organizations: software of the mind : intercultural cooperation and its importance for survival. (3rd ed) McGraw-Hill Professional ISBN 0071664181, 9780071664189

[24] Aaron Marcus and Associates, Inc . 2010. Cross-Cultural User-Interface Design for Work, Home, and On the Way: Accounting for Cultural Preferences, Acceptance, and Constraints. ACM SIGGRAPH ASIA 2010 Courses, Doi: 10.1145/1900520.1900525

[25] Rehm, M., Bee, N., and André, E. 2008. Wave like an Egyptian — Acceleration based gesture recognition for culture-specific interactions. In *Proceedings of HCI 2008*

Culture, Creativity, Interaction, 1, pp. 13–22, ISBN: 978-1-906124-04-5

[26] Wing Chee, S. 2010. "Cross-cultural transfer in gesture frequency in Chinese–English bilinguals". *Language and Cognitive Processes*, 25 (10), pp. 1335-1353, doi: 10.1080/01690961003694268

[27] Nicoladis, E. 2007. The effect of billingualiam on the use of manual gestures. *Applied Psycholinguistics*, 28, pp. 514-526.

[28] Yammiyavar, P., Clemmensen, T., and Kumar, J. (2008). Influence of cultural background on non-verbal communication in a usability testing situation. *International Journal of Design*, 2(2), pp. 31-40, http://www.ijdesign.org/ojs/index.php/IJDesign/article/view/313/164 [Accessed: December 15, 2012]

[29] Fikkert, W., Vna der Vet, P., Van der Ver, G., and Nijholt, A. 2009. Gestures for large display control . *Computer Interaction: 8th International Gesture Workshop*, Revised Selected Papers, pp. 245-256. Germany: Springer

[30] Hofstede, G. 2011. Dimensions of national Cultures., http://www.geerthofstede.nl/dimensions-of-national-cultures [Accessed: December 15, 2012]

[31] Hall, E. T. 1976. *Beyond Culture*. Knopf Doubleday Publishing Group. ISBN 0385124740, 9780385124744

[32] Liu, J. and Kavakli, M. 2011. Temporal Relation between speech and co-verbal iconic gestures in multimodal interface design, http://coral2.spectrum.uni-bielefeld.de/gespin2011/final/Liu.pdf [Accessed: December 15, 2012]

[33] Pause .2011. In *Oxford Dictionaries*, http://oxforddictionaries.com/definition/pause [Accessed: December 15, 2012]

Modeling in a Wiki with MoKi: Reference Architecture, Implementation, and Usages

Chiara Ghidini, Marco Rospocher, and Luciano Serafini
Fondazione Bruno Kessler - Centro per la Ricerca Scientifica e Tecnologica (FBK-irst),
Via Sommarive 18 Povo, I-38123, Trento, Italy
e-mail: {ghidini,rospocher,serafini}@fbk.eu

***Abstract*—The success of wikis for collaborative knowledge construction is triggering the development of a number of tools for collaborative conceptual modeling based on them. In this paper, we present a reference architecture for wiki-based collaborative conceptual modeling tools. The characteristics of our reference architecture are: (i) the use of wiki pages to describe semantic terms and organizational mechanisms of a conceptual modeling language; (ii) the organization of wiki pages in an unstructured part and a structured part; and (iii) a multi-mode access to the pages. We also present a completely revised version of the MoKi tool fully compliant with the presented reference architecture. A detailed description of some usages of MoKi in different application contexts is also reported.**

***Keywords-Conceptual Modeling*; *Collaborative Modeling*; *Semantic Wikis*; *Ontology Modeling*; *Process Modeling*.**

I. Introduction

In this paper, we present the wiki-based collaborative conceptual modeling tool MoKi, its reference conceptual architecture, the specific current implementation and usages in real-world scenarios. The paper extends the work presented at the "Fourth International Conference on Information, Process, and Knowledge Management (eKNOW 2012)" [1]. More in detail,

- the description of the reference conceptual architecture presented in Section III has been extended with a more detailed and formal description of how to represent conceptual modeling languages in a wiki, with the aim of providing a general conceptual architecture for collaborative conceptual modeling tools which encompasses MoKi;
- the description of the MoKi tool presented in Section V has been extended with more examples and the description of additional functionalities; and finally,
- an entire new section describing the usage of MoKi in different application contexts has been added (Section VI).

The idea of building wiki-based modeling tools has emerged in recent years as a promising approach towards the collaborative construction and visualization of conceptual models. In fact, from the success of Wikipedia onwards, wikis have been increasingly adopted as tools for collecting, sharing and managing knowledge, both in the case of domain specific knowledge (e.g., in enterprises) and in the case of encyclopedic knowledge. Moreover, recent projects such as DBpedia [2], YAGO [3], and Semantic Media Wiki (SMW) [4] have empowered traditional wikis with the capability of publishing their (usually) unstructured text and multimedia content in a structured, RDF-based, format. This has enabled users to employ better search, browse, and share facilities, and has extended the power of wikis transforming them from tools for the collaborative creation and management of content, to tools for the collaborative creation and management of (on-line) data and knowledge bases. This, in turn, has prompted the idea of building wiki-based tools for the collaborative construction and visualisation of conceptual models (see, e.g., the Halo extension and SMW+ [5], MoKi [6], and Ontowiki [7]), and has suggested the usage of the wiki philosophy in tools which are not directly built on top of wikis (e.g., Senso Comune [8], Freebase [9], and PoolParty [10]).

Despite this great amount of work, building a wiki-based tool for the modeling of a specific domain remains a challenging task, as the basic features of wikis must be used in a way that effectively support the construction of good quality conceptual models. The development of a clear reference architecture, where the focus is placed on identifying the key constructs and abstractions rather than on the technical characteristics of the tools themselves, would provide a significant contribution to meet this challenge. In this paper we address this task, taking into account the following needs:

- **Generality**. Until now, the work in the area of wiki-based modeling tools has mainly focused on the development of instruments targeted to specific conceptual models: thesauri, ontologies, RDF content, organizational / workflows, and so on. While this has contributed to show the potential of wikis, it has also delayed the emergence of a wiki-based paradigm for conceptual modeling. Defining a general paradigm for different conceptual modeling languages is a crucial step as it enables the use of similar abstractions and features for different types of models (e.g., an ontology or a workflow). This becomes especially important when users need to build, share, browse scenarios composed of different models. Think, for instance, to the case of an enterprise which needs to model an ontology

of competences together with the processes that need them, or to the case of a user who needs to browse a workflow together with the taxonomy used to annotate it. The reference architecture must aim at understanding how the features of wikis can be used to represent the building blocks of a general conceptual modeling language, before tailoring them to the needs of a particular one.

- **Collaboration**. A crucial step in building good quality conceptual models is the involvement of domain experts in the modeling process. As argued in [11], traditional methodologies and tools are based on the idea that knowledge engineers drive the modeling process. This often creates an extra layer of indirectness which makes the task of producing and revising conceptual models too rigid and complex, e.g., for the needs of business enterprises. In addition, the leading role of knowledge engineers can hamper the model construction as the domain experts (and domain knowledge) may become secondary to the process of efficient knowledge modeling, especially when domain experts have no understanding of the languages and tools used to build the conceptual models. The reference architecture must aim at understanding how the features of wikis can be used to support a well-balanced collaboration between domain experts and knowledge engineers in modeling process.

The contribution of this paper is manyfold. First, we present a reference architecture for wiki-based conceptual modeling tools which satisfies the two needs described above. The distinctive characteristics of our architecture are: (i) the use of wiki pages to mimic the basic building blocks of conceptual modeling languages, namely semantic terms and structuring mechanisms; (ii) the organization of wiki pages for semantic terms in an unstructured part (for unstructured content) and a structured part (for structured content); and (iii) a multi-mode access to the pages to facilitate the usage by domain experts and knowledge engineers. Second, we illustrate an implementation of this architecture in a completely revised version of MoKi [6]. This description aims at showing the feasibility of the proposed architecture by means of a practical realization. Third we report on the usage of MoKi in different application contexts. Again this description aims at showing the sustainability of the proposed architecture by means of concrete usage in real application contexts.

The novelty of our work can be found at different levels: at a *foundational* level, this paper provides the first architectural model for wiki-based conceptual modeling tools, which can be used to implement tools for different conceptual modeling languages in a uniform manner; at an *architectural* level, it introduces the idea of multi-mode access to pages to support easy usage both by domain experts and knowledge engineers; at the *implementation* level, MoKi provides the first attempt to build a single tool for different conceptual modeling languages able to support the collaboration of domain experts and knowledge engineers through the usage of a multi-mode access to knowledge.

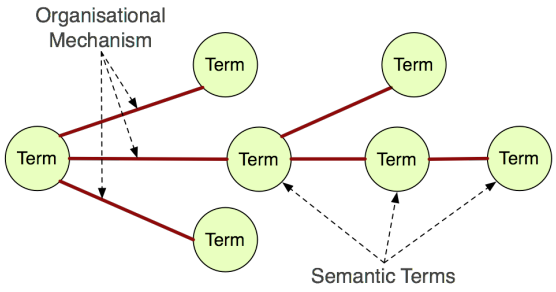


Figure 1. Conceptual Modeling Languages.

The paper is structured as follows: we start from an analysis of conceptual modeling languages (Section II) and we proceed by defining an architecture which satisfies the needs of generality and collaboration (Sections III and IV). We then provide a description of MoKi (Section V) and some application contexts in which it has been used (Section VI), concluding with a comparison between the proposed architecture and state of the art tools for wiki-based conceptual modeling (Section VII).

II. Conceptual Modeling

Conceptual modeling (aka semantic modeling) has been researched and used in several areas of Computer Science and Engineering, such as Database / Knowledge Representation, Software Engineering, and Artificial Intelligence, often with different usages, characterizations, and terminologies. According to [12] and [13], we can say that conceptual models provide a description of knowledge based on the so-called associationist viewpoint, where knowledge is organized in terms of: (i) nodes that represent concepts, and (ii) associations (or, links) that represent relationships between them. In particular, [13] provides a characterization of Conceptual Modeling Languages (CMLs) in terms of their two main building blocks, also illustrated in Figure 1:

1) **Semantic terms**: these are the concepts built into the conceptual model. Semantic terms are used to describe different types of concepts, such as Entities, Activities, Agents, Goals, and so on, depending on the CLM used; and
2) **Organizational mechanisms** (also called Abstraction Mechanisms in [13]): these are primitive mechanisms for structuring the model along different dimensions. Examples of abstraction mechanisms are: generalization (often referred to as *isA*), aggregation (*partOf*), classification (*instanceOf*), contextualization / modularization, and so on.

The different uses of Conceptual Models in the diverse areas of Computer Science and Engineering had important consequences on the development of specific CMLs. If the models are used mainly by people, e.g., to capture, organize and communicate high level knowledge, then the CML notation may be semi-formal or even informal, as in the case of Concept Maps, where no, or an extremely informal, semantics is usually associated to the diagrams. On the contrary, if the models need to be as less ambiguous as possible, and/or they need to be algorithmically exploited by computers to provide services such as consistency analysis or query answering, then the notation needs to correspond to a precise formal semantics, as in the case of OWL ontologies. In between these extreme cases are "semi-formal" CMLs. An example is the Business Process Modeling Notation [14], which provides a very detailed and specific syntactic notation with a semi-formal semantics.

III. Conceptual Modeling in Wiki Pages

The first challenge for wiki-based modeling tools is to be able to represent the two basic building blocks of conceptual modeling languages, namely semantic terms and organizational mechanisms. In this section we provide a definition of Conceptual Modeling Wiki (CMW) which uses the notion of wiki pages to represent these building blocks.

Let Σ be an alphabet in some CML $\mathcal{C}$, and Γ be the set of organizational mechanisms in $\mathcal{C}$. A *Conceptual Modeling Wiki (CMW)* $\mathcal{W}$ for Σ and Γ is a tuple $(P, SP, \mathsf{p}, \mathsf{a})$ where P is a non empty set of wiki pages, SP is a non empty set of special pages, $\mathsf{p} : \Sigma \to P$, and $\mathsf{a} : \Gamma \to SP$. We say that:

- $P \cup SP$ is the set of pages of the wiki $\mathcal{W}$;
- $\mathsf{p}(\alpha)$ is the wiki page associated to the semantic term $\alpha \in \Sigma$;
- $\mathsf{a}(\beta)$ is the set of (one or more) wiki special pages associated to the organizational mechanism $\beta \in \Gamma$.

A pictorial representation of a conceptual modeling wiki is given in Figure 2. In a nutshell, a wiki is composed of a set $P \cup SP$ of pages, where each (regular) page in P is used to describe semantic terms in the model, and each special page in SP is used to display a functionality which enables the browsing / editing of the overall organization of the conceptual model according to a specific organizational mechanism. For instance, if we consider a CML which contains semantic terms for concepts, instances, and roles, and two organizational mechanisms such as generalization and aggregation, then we need a wiki able to associate a regular wiki page to each semantic term of type concept, instance, and role, plus two special pages which enable to visualize (edit) the overall model organised according to the generalization and the aggregation/decomposition dimensions respectively.

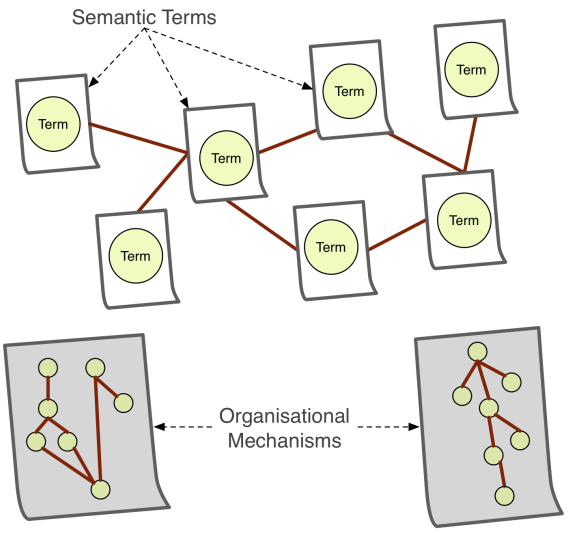


Figure 2. Representing a conceptual model in a wiki.

A. Building Wiki Pages for Terms

The idea of associating a wiki page to each semantic term, contained in the definition of Conceptual Modeling Wiki, is adopted by most of the state of the art wiki-based tools used to represent and manage knowledge (see Section VII). Nevertheless, this first idea needs to be refined and expanded if we aim at providing tools able to exploit in full the wiki potential and to make all the actors of the modeling team collaborate towards the creation, modification and exploitation of knowledge.

An important characteristic of wiki-based tools is their capability to deal with both *structured* and *unstructured* content. Assume, for instance, that we have to describe the term "Mountain". We can describe it in a "wikipedia style", by using text and pictures, as for instance is done in the "Mountain" page in Wikipedia [15], or we can provide more structured descriptions, in the style of Freebase, Ontowiki or of a Wikipedia Infobox. In this paper we argue that both types of content are essential in a process of conceptual modeling, and that a wiki page for a semantic term should be composed of two parts: the *unstructured part* and the *structured part*, as depicted in Figure 3. The first, unstructured part contains the rich and often exhaustive descriptions of knowledge which is better suited to humans and is built using linguistic and pictorial instruments. While some guidelines can be provided to organize the unstructured part, asking for instance for definitions, descriptions of the main characteristics, samples individuals (prototypes), a gallery of pictures, related/relevant documents, and so on, the content of this part of the page has a high degree of freedom. The second, structured part is instead the one which is used to provide the portion of knowledge which will be directly encoded in the CML. Differently from the unstructured part,

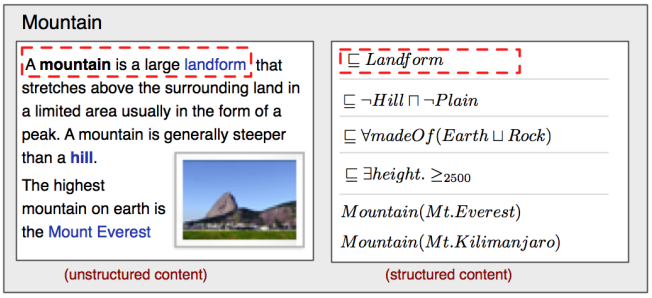


Figure 3. Wiki page for semantic terms.

which is expressed using natural language and multimedia content, the structured part of the page can have different formats, according to the CML used. Examples are: simple statements which describe the attributes of the semantic term being described; a list of inclusions axioms defining a concept in OWL (as in Figure 3); diagrams expressed in a workflow (business process) oriented language, and so on.

More formally, let $\alpha \in \Sigma$ be a semantic term of some CML $\mathcal{C}$. We propose that a *page* $\mathsf{p}(\alpha)$ for α in a wiki $\mathcal{W}$ is a pair (u,s), where u is a regular wiki string, that possibly contains links to other wiki pages of the wiki $\mathcal{W}$, and s is a description of that term in $\mathcal{C}$ using the alphabet Σ.

The advantage of storing the unstructured and structured descriptions in the same tool is twofold. First, the informal descriptions are usually used both to provide the initial description upon which the formal model is built, and to document the elements of the model, e.g., for future access and revisions. Storing the unstructured and structured descriptions in the same tool can facilitate the interplay between these parts, e.g., by adding alignment functionalities. Second, domain experts, who usually create, describe, and review knowledge at a rather informal/human intelligible level, may find the unstructured part their preferred portion of page where to describe knowledge. Instead, knowledge engineers should be mainly focused on the descriptions contained in the structured part. Nevertheless, by using the same tool and accessing the same pages they can be notified of what the others are focused at. Moreover, the discussion facilities of wikis, together with special fields for notes and comments, can be used by both roles to discuss and collaborate on specific parts of the model.

Note that, while a complete alignment between the unstructured and structured parts of a wiki page is not achievable, and most likely not even appropriate, as the rich nature of the unstructured representation is often not meant to be entirely transferred in a formal representation, it is easy to observe that specific portions of the unstructured part can provide descriptions upon which a certain piece of the structured representation is based, or can provide documentation which justifies or explains parts of the structured description (see, e.g., the two sentences surrounded by dotted lines in Figure 3). Manual or semi-automatic functionalities to interlink the content contained in the unstructured and structured descriptions should therefore be provided in a CMW to support the interplay between the unstructured and structured knowledge contained in the wiki.

IV. Supporting Multi-mode Access to Conceptual Models

The organization of a page in an unstructured and structured part is a second important step in defining the architecture of a conceptual modeling wiki, but may not be enough in the case of complex CMLs, such as the ones based on logical formalisms (e.g., OWL [16]) or very complex notations (e.g., BPMN [14]). In this case the structured part of the page will contain very precise, and often logic based, descriptions of a term, preventing domain experts from accessing the domain knowledge encoded in the conceptual model.

To overcome this problem, we propose to separate the content of the page from the functionalities used to view and edit it. Hereafter we call these functionalities *access modes*. The idea of this novel characteristic of wiki-based tools for conceptual modeling is to associate different access modes to each part of the page, as depicted in Figure 4, to enable a multi-mode access to the content stored in the page. In the example of the wiki page for "Mountain", introduced in the previous section and depicted in Figure 4, the unstructured content is stored in a regular wiki string and the structured content is stored in OWL. Therefore, the access mode to the unstructured part can be provided by means of the regular view/edit facilities of wikis, while the access to the structured content can be provided by means of two different modes: one based on a translation of the OWL content in, e.g., DL axioms or in the Manchester OWL syntax, and another based on a structured, but semi-formal rendering of the OWL content in a pre-defined template as the one depicted at the bottom of Figure 4. In this way the knowledge engineers can formally describe the semantic term "Mountain" in the chosen CML by using a highly formal access mode, while the domain experts can access a simplified version of the same content using a different, simpler, mode.

More formally, the *multi-mode access* of a page (u,s) for a semantic term α is a pair $(\{uam_1,\ldots uam_n\},\{sam_1,\ldots,sam_k\})$, $n,k \geq 1$, where each uam_i is an access mode for u and each sam_j is an access mode for s.

As we can see from the definition above we can potentially define a number of different access modes for each part of the page, which can be based on the different existing approaches towards representation of content and knowledge. Examples are: different access modes which represent the OWL structured content using different syntax, controlled natural languages, or graphical representations. Analogously

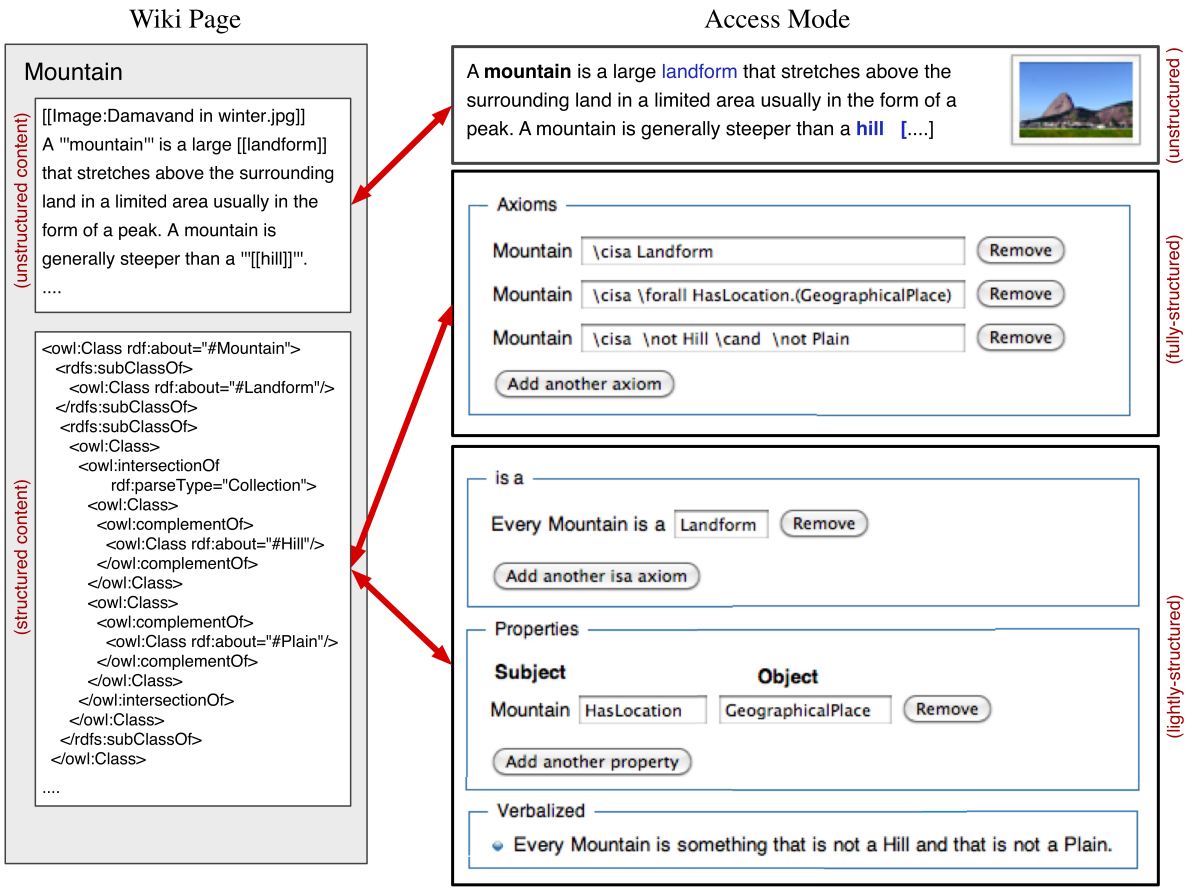


Figure 4. Multi-mode access to a wiki page for semantic terms.

we can have different templates which render the structured content at a different levels of complexity. Nevertheless, we believe that CMW tools for highly structured CMLs should be based on (at least) three different access modes:

- a *unstructured access mode* to view/edit the unstructured content;
- a *fully-structured access mode* to view/edit the complete structured content; and
- a *lightly-structured access mode* to view/edit (part of) the structured content via simple templates.

We propose these three modes only for highly structured CMLs as the distinction between fully-structured and lightly-structured access modes may become unclear in case of simple CMLs with informal semantics such as concept maps. In these cases the fully structured representation is often simple enough to be directly accessible also by domain experts.

The advantage of providing two distinct modalities to access the structured content of a wiki page lies in the ability of providing an access to the conceptual model to both domain experts and knowledge engineers. In this way domain experts can not only have access to the knowledge inserted by knowledge engineers, but can also comment or directly modify part of it. An important aspect of the implementation of a CMW is therefore the design of appropriate access modes, which can be based on templates whose formats depend upon the CML used and also upon the degree of complexity handled by the domain experts. Examples of templates which can be used to provide a lightly-structured access mode are: (possibly simplified) verbalizations of OWL statements; simple flow diagrams which represent the main steps of a workflow (business process); matrixes which provide a diagrammatic representation of binary roles; and so on. Another important aspect in the implementation of a CMW is the interaction between the structured content *s* and the lightly-structured access mode. Differently from the unstructured access mode and fully-structured access mode where the content shown/edited within the access mode can be considered a one-to-one syntactic variant of the content stored in the page, this is not the case for the lightly-structured access mode. In fact, the content stored in the structured part may be too expressive or complex to be directly represented in the lightly-structured access mode. In

this case, functionalities must be provided to "translate" the structured content of the page in the simplified representation in the lightly-structured access mode, and vice-versa.

V. Conceptual modeling with MoKi

MoKi [17] is a collaborative MediaWiki-based [18] tool for modeling ontological and procedural knowledge in an integrated manner. MoKi uses OWL (Description Logics) and BPMN as the reference CMLs for ontological and procedural knowledge respectively, and associates semantic terms of the two CMLs to wiki pages containing both unstructured and structured information, accessible using different access modes.

In this section, we present a completely revised version of MoKi, which extends the first release of the tool (see [6]) to be fully compliant with the architecture illustrated in Sections III–IV. The main changes w.r.t. [6] are:

(i) a redesign of the content organization of the MoKi page, which now comprises an unstructured part and a structured part. This extends and replaces the simple representational languages used in [6], and enables to model rich semantic terms using expressive ontology and complex business process CMLs; and

(ii) a new support for multi-mode access to the page content which implements the three different access modes described in Section IV. This extends and replaces the single template-based access mode provided in [6].

A. The MoKi page for a semantic term

Being a tool supporting the description of ontological and procedural knowledge according to OWL and BPMN, the types of semantic terms relevant for MoKi are *concepts*, *properties*, and *individuals* in the ontology, and *process* (in MoKi we use the term "process" as a synonym for (complex or simple) activity) in the process model. Each term belonging to one of these types is therefore associated to a MoKi page which, coherently with the discussion in Section III-A, is composed of an unstructured part and a structured part.

The unstructured part: This part contains text written following the standard MediaWiki markup format: in particular, it can contain plain text, possibly enriched by formatting information, links to other MoKi pages or to external resources, uploaded images, and so on. The format of this part of the page is the same for all the different semantic terms.

The structured part: This part, which is delimited by specific tags to separate it from the unstructured text, contains knowledge stored according to the CML adopted. In the current implementation, the structured part of a page describing an ontology term contains a RDF/XML serialization of a set of OWL statements formalising the term, while, similarly, the structured part of a page describing a BPMN process contains a serialization of the process diagram in the JavaScript Object Notation (JSON).

B. Supporting multi-mode access in MoKi

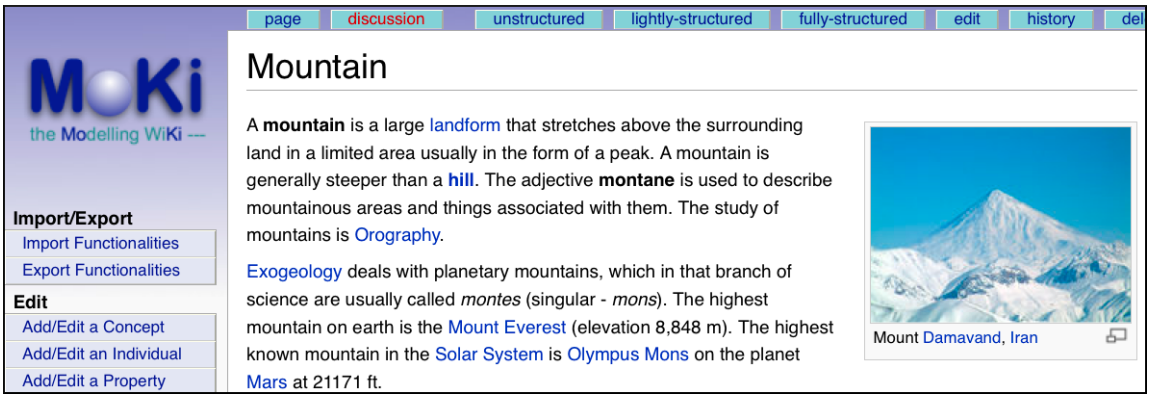
Users can access the ontological and procedural knowledge contained in MoKi using the three different access modes described in Section IV: one mode, the *unstructured access mode*, to access the unstructured part of a MoKi page, and two different modes, the *fully-structured access mode* and the *lightly-structured access mode*, to access the structured part.

The unstructured access mode: This access mode allows the user to edit/view the content of the unstructured part of the MoKi page of a semantic term. The editing/viewing of this part occurs in the standard MediaWiki way. Figure 5(a) shows the unstructured access mode of a portion of MoKi page describing the concept of "Mountain" (the content of the page in the figure is an excerpt taken from Wikipedia [15]).

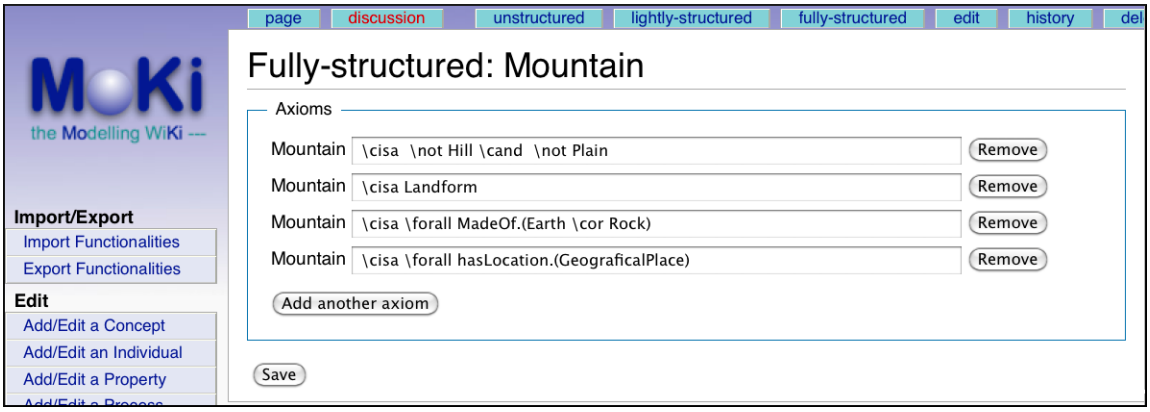
The fully-structured access mode: This access mode allows the user to edit/view the content of the structured part of a MoKi page using the full expressivity of the chosen CML. For ontological knowledge the fully-structured access mode allows the user to view/edit formal statements (axioms) describing the term associated to the page. Axioms are written according to the *latex2owl* syntax [19], an intuitive latex-style format for writing ontologies using a text-editor, format which can be automatically translated into (an RDF/XML serialization of) OWL. The latex2owl syntax was chosen because of its resemblance to the DL syntax; however, the approach illustrated here can be easily used to support a fully-structured access mode based on other OWL syntaxes such as the Manchester OWL syntax [20]. The user can easily edit the list of axioms in a form based interface, as the one shown in Figure 5(b). When saving the page, all axioms in the page are translated in OWL by the latex2owl tool, and the resulting code is stored in the structured part of the page. Conversely, when loading the page, the owl2latex tool translates the OWL code into statements adherent to the latex2owl syntax.

For procedural knowledge we have implemented an access mode that allows the user to edit the BPMN process diagram described in the page as shown in Figure 6(a). In particular we have tightly integrated in MoKi the Oryx editor [21], a full-fledged business process editor that allows to create processes according to several modeling languages, including BPMN.

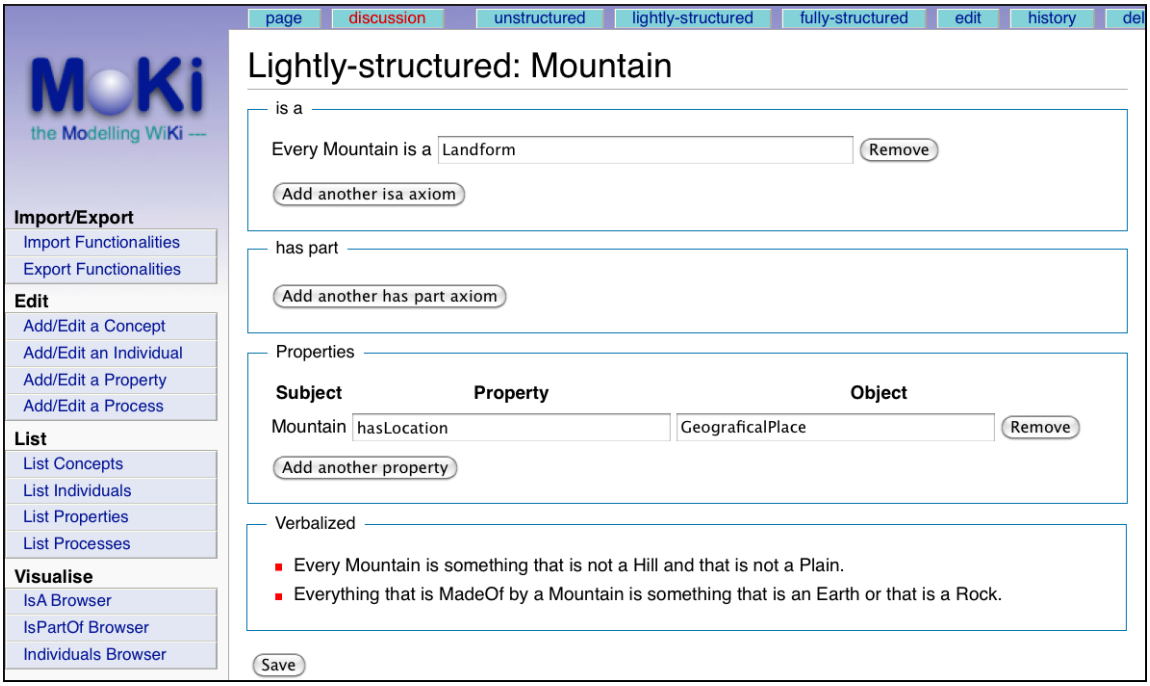
The lightly-structured access mode: As described in Section IV the purpose of this access mode is to allow users with limited knowledge engineering skills, to edit/view the content of the structured part of the MoKi page in a simplified and less formal way. For ontological knowledge the lightly-structured access mode is provided through a form made of two components, as depicted in Figure 5(c). In the top half part the user can view and edit simple statements which can be easily converted to/from OWL statements. For instance, in the case of concepts the user can edit statements



(a) Unstructured access mode.

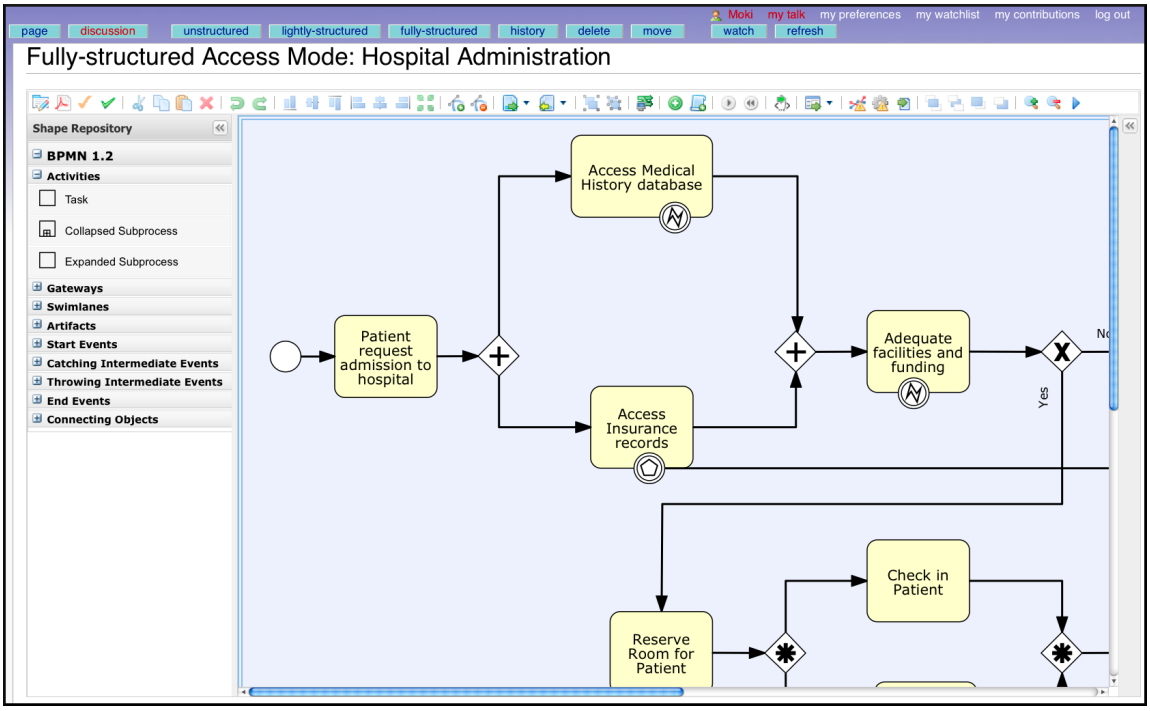


(b) Fully-structured access mode.

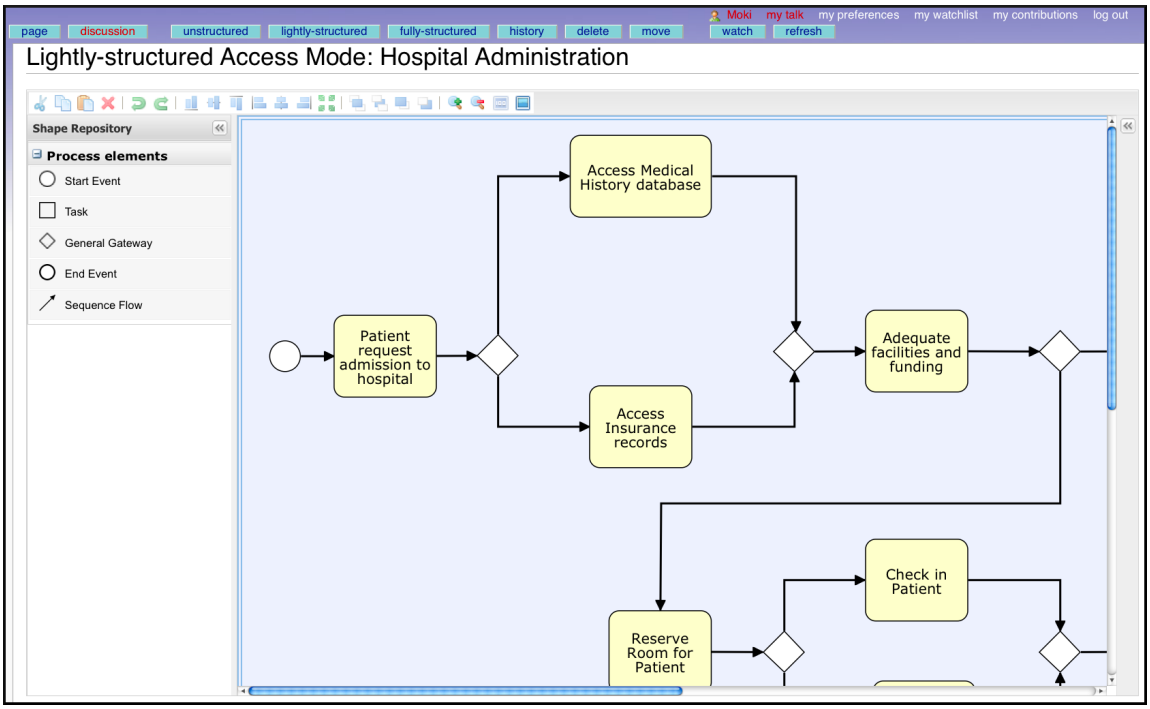


(c) Lightly-structured access mode.

Figure 5. Multi-mode access to the page of concept *Mountain*.



(a) Fully-structured access mode.



(b) Lightly-structured access mode.

Figure 6. Fully-structured access mode and lightly-structured access mode to the page of a process about Hospital Administration.

of the form "Every *subject* is a *object*", "Every *subject* has as part a *object*", or, more generally, statement of the forms (*subject, property, object*), which correspond to the latex2owl statements "*subject* \cisa *object*", "*subject* \cisa \exists hasPart.*object*", and "*subject* \cisa \forall *property*.(*object*)" respectively. Analogous forms are provided

for properties and individuals. If the OWL version of any of these statements is already contained in the structured part of the page, then the corresponding fields are pre-filled with the appropriate content. Similarly, when any of these simple statements is modified in the lightly-structured access mode, the changes are propagated to the content of the structural part of the page. The bottom half of the form provides a description of those OWL statements which cannot be intuitively translated/edited as simple statements as the ones in the top half of the page. In the current implementation, this part contains the translation of those statements in Attempto Controlled English, provided by the OWL 2 Verbalizer [22]. The purpose of this bottom half of the form is to give the domain experts a flavour of the complex statements that a knowledge engineer has formalized. If a domain expert is doubtful about some of the statements, he/she can mark them and ask for a clarification using, e.g., the MediaWiki Discussion functionality.

For procedural knowledge we have implemented an access mode based on a light-weight graphical process editor such as the one shown in Figure 6(b). This editor shows only the basic workflow of the activity, and the main elements of the process such as start and end events, plus the sub-processes it can contain, hiding the details and complexity typical of BPMN diagrams. The challenges we are currently addressing are how to visualise in a simplified way a complex BPMN process stored in the structured part, and how to update the content of the structured part according to the changes performed in the lightly-structured access mode.

C. Organizational mechanism pages in MoKi

Organizational mechanism pages are MoKi special pages dynamically created from the (structured) content of the semantic term pages. Differently from wiki pages for terms, which are mainly constructed using textual representations, the organizational mechanism rely also on graphical forms of representation, which include graphical browsing and editing facilities. For ontological knowledge the organizational mechanism pages allow to explore the generalization and part/subparts decomposition hierarchies of ontology concepts, as well as the classification of the ontology individuals. In particular, MoKi provides two kinds of organizational mechanism pages. In the tabular-based one, the user can access a table listing every concept (resp. individual) of the ontology together with the concepts of which it is a specialization and the concepts in which it decomposes according to the part of relation (resp. the concepts to which the individual belongs to). In the graphical-based one, a tree-like view shows the hierarchy of concepts according to either the subclass (see Figure 7) or the part-of relation, or the membership of individuals to concepts. Drag and drop editing facilities are also provided to rearrange the tree. For procedural knowledge, the current organizational mechanism page provides an overview of the activity/sub-activity decomposition mechanism by means of a tree-based view as well as a table listing every process defined in MoKi together with the processes in which it decomposes. Already planned work aims at including a workflow-based representation of the before/after abstraction mechanism, which, in the current version, is limited to the description of the sub-process which represent how a complex activity is structured, as depicted in Figure 6.

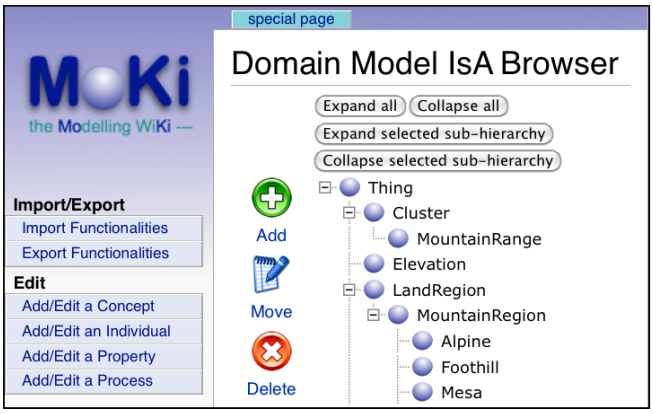


Figure 7. Example of organizational mechanism (generalization) page: the IsA Browser.

VI. Usages of MoKi

MoKi has been successfully applied in several application contexts. Next, we report some of them, describing the specific tasks for which MoKi was used, as well as some of the findings and insights we observed. Detailed users experiments and evaluations are reported in [23], [24].

A. APOSDLE Project

The APOSDLE EU project [25] developed a software platform and tools to support the process of *learning@work*, that is learning within the context of the immediate work of a user and within her current work environment. To deliver a user with context-sensitive learning material or suggestions, tailored to her specific needs, the APOSDLE system needs to know not only the profile of the user, but also about various aspects of the sphere in which the user is acting, spanning from the specific domain of knowledge in which the user is working, the tasks (aka processes) the user can perform, the learning goals she can have, and also the material available to compose adequate learning suggestions.

An early version of MoKi [26] have been successfully applied to develop integrated models of ontology and process for the purpose of initialising and serving as the knowledge back-end of the APOSDLE platform. Six different domains were considered: Information and Consulting on Industrial Property Rights (94 concepts and 2 properties; 13 processes), Electromagnetism Simulation (115 concepts and 21 properties; 13 processes), Innovation and Knowledge Management

(146 domain concepts and 5 properties; 31 processes), Requirements Engineering (the RESCUE methodology) (78 concepts and 2 properties; 77 processes), Statistical Data Analysis (69 concepts and 2 properties; 10 processes) and Information Technology Infrastructure Library (100 concepts and 2 properties; no processes).

The modelling activities involved people with different modelling skills and levels of expertise of the application domains, and located in different places all over Europe. The evaluation of the version of MoKi used in the project showed that the users highly appreciated the form-based interface of MoKi, and the fact that they were able to participate in the creation of the models without having to know any particular syntax or deep knowledge engineering skills. Thus, MoKi was perceived as an adequate tool to actively involve domain experts in the modelling process. People with some knowledge engineering skills found MoKi as comfortable to use as other state-of-the-art modelling tools, and MoKi helped the users in structuring and formalizing their knowledge in a simple, intuitive and efficient manner. Particularly appreciated have been the functionalities, in particular the graphical ones, which allow to browse the models along the organizational abstractions.

B. PESCaDO Project

Citizens are increasingly aware of the influence of environmental and meteorological conditions on the quality of their life. One of the consequences of this awareness is the demand for high quality environmental information and decision support that is tailored (i.e., personalized) to ones specific context and background (e.g., health conditions, travel preferences). Personalized environmental information may need to cover a variety of aspects (e.g., meteorology, air quality, pollen) and take into account a number of specific personal attributes of the user (e.g., health, age, allergies), as well as the intended use of the information. The goal of the PESCaDO EU project [27] is to develop a multilingual web-service platform providing personalized environmental information and decision support.

The backbone of the PESCaDO platform [28] is an environmental ontology-based knowledge base where all the information relevant for a user request are dynamically instantiated. The ontology formalizes a variety of aspects related to the application context: environmental data, data sources, user requests, user profiles, warnings and recommendations triggered by environmental conditions, and so on. MoKi is being exploited by some knowledge engineers in the consortium to support the construction and revision of this ontology, also by exploiting some automatic ontology concepts extraction techniques offered by the tool [29], [30] (functionalities that may also be exploited for terminologically evaluating ontologies [31]). The current version of the ontology consists of 241 concepts, 151 object properties, 43 datatype properties, and 672 individuals.

C. Organic.Edunet and Organic.Lingua Projects

Organic.Edunet [32] was a EU project that aimed to facilitate access, usage and exploitation of digital educational content related to Organic Agriculture (OA) and Agroecology. It deployed a multilingual online federation of learning repositories, populated with quality content from various content producers. In addition, it deployed a multilingual online environment (the Organic.Edunet portal) that facilitate end-users search, retrieval, access and use of the content in the learning repositories.

MoKi was used by a team of knowledge engineers and domain experts to collaboratively build and revise the Organic Agriculture and Agroecology Ontology (61 concepts, 30 properties, and 222 individuals) at the core of the Organic.Edunet portal. The experience was perceived as positive enough by the user to favour the adoption of MoKi as the central modelling tool in the follow-up project, Organic.Lingua.

Organic.Lingua [33] is a EU project aiming to enhance the existing Organic.Edunet Web portal with educational content on Organic Agriculture (OA) and Agroecology (AE), introducing automated multi-lingual services that will further support the uptake of the portal from its targeted audiences, facilitate the multilingual features of the portal, and further extend its geographical and linguistic coverage.

The version of MoKi that has been customized for Organic.Lingua addresses the challenges posed by the project, that is to manage the multilingual aspects of the ontology used to tag the resources deployed on the portal, and the internationalization of the tool itself. Therefore, the Organic.Lingua MoKi implements features that permit to manage: the translations of each ontology entity name and description, the discussions about the changes that have to be carried out on the ontology, and the translations of the interface labels.

D. ProDe Project

ProDe [34] is an Italian inter-regional project with the aim of defining a national reference model for the management of electronic documentation (dematerialized document) in the Public Administration. This reference model follows an archival science perspective, and can be used for the identification of guidelines and functions needed to safely store, classify, manage, and retrieve, electronic documents produced within the PA in an archival system.

A customized version of MoKi [35] supported teams of users (both domain experts and knowledge engineers involved in the project) in the construction of the reference model, which consisted of an *ontological* part, formalizing document management and organizational aspects, i.e., document archiving-related aspects and the offices and profiles involved in the dematerialization, as well as a *process* part, describing the activities which produce (manage, consume) documents.

A quantitative and qualitative evaluation of the usage of MoKi in ProDe was performed [24]. Users perceived the tool as more than easy to use, and positively rated the overall usefulness of the tool for the collaborative modeling of documents and processes.

E. ProMo Project

ProMo [36] is an industrial project (FESR) founded by the province of Trento that aims at the development of a platform that supports the collaborative modelling of the processes, the structure, the actors, and the artifacts of a complex organization like a Public Administration, the grounding of the objects in this abstract model to the technological layer of the organization (services execution platform), and the support for monitoring the process execution.

MoKi is currently used by a team knowledge engineers to formally describe some exemplar processes of the Public Administration, as for instance the procedure for registering new citizens in a town hall. These processes are enriched with entities defined in a domain ontology, defined in MoKi as well, that provides the domain knowledge semantic of the processes elements, in line with the semantic annotation of business processes approach presented in [37].

F. OncoCure and eOnco Projects

The general aim of the OncoCure project [38], [39] was to use innovative ICT-based methods and models for clinical governance in oncology, by designing and developing a system for supporting and controlling the best evidence-based oncological care process. The system is based on electronic guidelines and recommendations, to be integrated with clinical information systems that manage the oncological patients. The main goal was the design and development of a prescriptive decision support system (DSS) for clinicians during the care process, based on the execution of AIOM (Associazione Italiana di Oncologia Medica) guidelines for breast cancer formalized in the Asbru, a plan-specification language for defining clinical protocols.

Although MoKi as presented here is tailored to the development of ontologies and business processes, a preliminary and customized version of the tool that supports the modelling of clinical protocols in the ASBRU modelling language is described in [40]. This version of MoKi, called *CliP*-MoKi, provides support for modelling the key elements of an Asbru model (e.g plans, parameters, http://www.asbrusoft.com/) as wiki pages, and for exploring the models created according to the organizational mechanisms for structuring knowledge provided by the language (e.g. the plan/plan children decomposition).

The eOnco project [41] had the main objective of supporting knowledge intensive management of cure process in Oncology. One of the aspects of the project concerned the elicitation and formal representation of the activities done by the nurses in a ward, to investigate for instance the bottlenecks, performance-wise, of the nurses work.

A multi-disciplinary team composed of sociologists and knowledge engineers participated to the formalization of these activities in MoKi, which resulted in 10 BPMN processes consisting of 140 activities.

VII. Related Work

To the best of our knowledge, there are no works in the literature that explicitly address the problem of defining a reference architectural model for wiki-based conceptual modeling tools. Partially related to the content of this paper is the work in [42], which investigates *concept* modeling approaches applied in semantic wikis, that is design patterns to organize data in wikis (e.g. RDF Modeling, Relational Modeling, Rule Modeling).

Focusing on tools, wiki systems and semantic wikis have been mainly applied to support collaborative creation and sharing of ontological knowledge. *AceWiki* [43] was developed in the context of logic verbalization, that is, the effort to verbalise formal logic statements into English statements and vice-versa. AceWiki is based on Attempto Controlled English (ACE), which allows users expressing their knowledge in near natural language (i.e. natural language with some restrictions). *Semantic MediaWiki+* [5], which includes the Halo Extension, is a further extension on Semantic MediaWiki with a focus on enhanced usability for semantic features. Especially, it supports the annotation of whole pages and parts of text, and offers "knowledge gardening" functionalities, that is maintenance scripts at the semantic level, with the aim to detect inconsistent annotations, near-duplicate entries etc. *IkeWiki* [44] supports the semantic annotation of pages and semantic links between pages. Annotations are used for context-specific presentation of pages, advanced querying, consistency verification or drawing conclusions. *OntoWiki* [7] seems to focus slightly more directly on the creation of a semantic knowledge base, and offers widgets to edit/author single elements/pages and whole statements (subject, predicate, object). Finally, a proposal of modeling straightforward workflows using Semantic MediaWiki is implemented in the *Semantic Result Formats* extension [45].

We have compared the tools mentioned above, plus the previous and current versions of MoKi, against the distinctive characteristics of our reference architecture. The results are displayed in the Table I. The columns of the table refer to the capability of: (i) associating a page to a semantic term (*one page/one term*); (ii) browsing / overviewing the model according to the some organizational mechanism (*overview*); (iii) describing a semantic term using both unstructured and structured content (*unstructured/structured*); (iv) accessing content in a multi-mode manner (*multi-mode*); and (v) defining models according to two or more (substantially different) CMLs (*multiple CMLs*).

Table I
COMPARISON OF STATE-OF-THE-ART MODELLING WIKIS.

	one page/ one term	overview	unstructured/ structured	multi-mode	multiple CMLs
AceWiki	✓				
SMW+	✓	✓	✓		
IkeWiki	✓		✓		
OntoWiki	✓	✓	✓		
Semantic Result Formats	✓	✓	✓		
MoKi v.1	✓	✓	✓		✓
MoKi v.2	✓	✓	✓	✓	✓

As we can see from Table I, the proposed architectural model takes into account typical characteristics of wiki based-tools for conceptual modeling, pointed out by the first three columns of the table, and enriches them with two novel aspects, namely the multi-mode access to pages and the focus on multiple CLMs.

VIII. CONCLUSIONS

In this paper, we have presented a reference architectural model for wiki-based conceptual modeling tools grounded on three distinctive characteristics; (i) the use of wiki pages to mimic the basic building blocks of conceptual modeling languages; (ii) the structuring of wiki pages for semantic terms in an unstructured part and a structured part; and (iii) a multi-mode access to the pages to support easy usage both by domain experts and knowledge engineers. We have also described a fully revised version of MoKi which complies with the proposed architectural model. Several application contexts in which MoKi has been successfully applied have also been presented.

In our future work, we aim at improving the support for process modeling, in particular in providing an extensive automatic support for aligning the fully-structured access mode and lightly-structured access mode for procedural knowledge. One of the key aspects on which we are currently working is on enhancing the support for collaboration between people who model at different levels of abstraction: in particular, we are implementing facilities to highlight changes across the different access modes, to make domain experts aware of the changes introduced by knowledge engineers and vice-versa.

ACKNOWLEDGEMENTS

APOSDLE was funded under the FP6 ICT Support Program of the EU Commission (Grant Agreement Number FP6-IST-2004-027023). PESCaDO is funded under the ICT Support Program of the EU Commission (Grant Agreement Number FP7-248594). Organic.Lingua is funded under the FP7 ICT Support Program of the EU Commission (Grant Agreement Number 270999). ProDe is an Italian interregional project. OncoCure was funded by the Fondazione Caritro of Trento. ProMo is funded by the Operational Programme FESR 2007-2013 of the Province of Trento.

REFERENCES

[1] C. Ghidini, M. Rospocher, and L. Serafini, "Conceptual modeling in wikis: a reference architecture and a tool," in *The Fourth International Conference on Information, Process, and Knowledge Management (eKNOW2012), Valencia, Spain*, 2012, pp. 128–135. [Online]. Available: http://www.thinkmind.org/download.php?articleid=eknow_2012_6_10_60015

[2] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "Dbpedia - a crystallization point for the web of data," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 7, no. 3, pp. 154–165, 2009.

[3] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago - a core of semantic knowledge," in *16th international World Wide Web conference (WWW 2007)*, 2007.

[4] M. Krötzsch, D. Vrandecic, M. Völkel, H. Haller, and R. Studer, "Semantic wikipedia," *Journal of Web Semantics*, vol. 5, pp. 251–261, 2007.

[5] D. Hansch and H.-P. Schnurr, "Practical applications of semantic mediawiki in commercial environments - case study: semantic-based project management," in *3rd European Semantic Technology Conference (ESTC2009)*, 2009.

[6] C. Ghidini, B. Kump, S. Lindstaedt, N. Mahbub, V. Pammer, M. Rospocher, and L. Serafini, "MoKi: The Enterprise Modelling Wiki," in *Proceedings of ESWC 2009*, ser. LNCS, vol. 5554. Springer, 2009, pp. 831–835, demo Session.

[7] S. Auer, S. Dietzold, and T. Riechert, "Ontowiki - a tool for social, semantic collaboration," in *Proceedings of the 5th International Semantic Web Conference, Nov 5th-9th, Athens, GA, USA*, vol. 4273. Springer, 2006, pp. 736–749.

[8] A. Oltramari and G. Vetere, "Lexicon and ontology interplay in senso comune," in *Proceedings of OntoLex 2008 (Hosted by Sixth international conference on Language Resources and Evaluation)*, Marrakech (Morocco), 2008.

[9] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proc, of the 2008 ACM SIGMOD international conference on Management of data*. New York,: ACM, 2008, pp. 1247–1250. [Online]. Available: http://ids.snu.ac.kr/w/images/9/98/SC17.pdf

[10] T. Schandl and A. Blumauer, "Poolparty: Skos thesaurus management utilizing linked data," in *The Semantic Web: Research and Applications*, ser. LNCS, vol. 6089, 2010, pp. 421–425.

[11] V. Dimitrova, R. Denaux, G. Hart, C. Dolbear, I. Holt, and A. G. Cohn, "Involving domain experts in authoring owl ontologies," in *Proceedings of the 7th Int. Semantic Web Conference (ISWC 2008)*, ser. LNCS, vol. 5318/2010. Springer Berlin / Heidelberg, 2008, pp. 1–16.

[12] M. Hammer and D. McLeod, "Database description with sdm: A semantic database model," *ACM Trans. Database Syst.*, vol. 6, no. 3, pp. 351–386, 1981.

[13] J. Mylopoulos, "Information modeling in the time of the revolution," *Information Systems*, vol. 23, no. 3-4, June 1998.

[14] OMG, "Business process modeling notation, v1.1," www.omg.org/spec/BPMN/1.1/PDF.

[15] Wikipedia, "Mountain — Wikipedia, the free encyclopedia," 2012, [Online; accessed 09-July-2012]. [Online]. Available: http://en.wikipedia.org/w/index.php?title=Mountain&oldid=500787709

[16] M. K. Smith, C. Welty, and D. L. McGuinness, "Owl web ontology language guide," W3C Recommendation, 10 February 2004.

[17] DKM Unit, "MoKi: the MOdelling wiKI," [Online; accessed 09-July-2012]. [Online]. Available: https://moki.fbk.eu/

[18] Wikimedia Foundation, "Mediawiki," http://www.mediawiki.org.

[19] D. Unit, "Latex2owl," 2012, [Online; accessed 09-July-2012]. [Online]. Available: http://dkm.fbk.eu/index.php/Latex2owl

[20] CO-ODE, "The manchester owl syntax," 2012, [Online; accessed 09-July-2012]. [Online]. Available: http://www.co-ode.org/resources/reference/manchester_syntax/

[21] University of Potsdam, "Oryx - web-based graphical business process editor," 2012, [Online; accessed 09-July-2012]. [Online]. Available: http://code.google.com/p/oryx-editor/

[22] K. Kaljurand and N. Fuchs, "Verbalizing OWL in Attempto Controlled English," in *Proceedings of Third International Workshop on OWL: Experiences and Directions (OWLED 2007)*, 2007.

[23] C. Di Francescomarino, C. Ghidini, and M. Rospocher, "Evaluating wiki-enhanced ontology authoring," in *The 18th International Conference on Knowledge Engineering and Knowledge Management (EKAW2012), Galway, Ireland*, 2012.

[24] C. Casagni, C. Di Francescomarino, M. Dragoni, L. Fiorentini, L. Franci, M. Gerosa, C. Ghidini, F. Rizzoli, M. Rospocher, A. Rovella, L. Serafini, S. Sparaco, and A. Tabarroni, "Wiki-based conceptual modeling: An experience with the public administration," in *The Semantic Web – ISWC 2011*, ser. LNCS. Springer, 2011, vol. 7032, pp. 17–32.

[25] APOSDLE Consortium, "Advanced Process-Oriented Self-Directed Learning Environment (APOSDLE)," 2006-2010, [Online; accessed 09-July-2012]. [Online]. Available: http://www.aposdle.org/

[26] C. Ghidini, M. Rospocher, L. Serafini, B. Kump, V. Pammer, A. Faatz, A. Zinnen, J. Guss, and S. Lindstaedt, "Collaborative knowledge engineering via semantic mediawiki," in *International Conference on Semantic Systems (I-SEMANTICS '08)*, S. Auer, S. Schaffert, and T. Pellegrini, Eds., Graz, Austria, september 3–5 2008, pp. 134–141.

[27] PESCaDO Consortium, "Personalized Environmental Service Configuration and Delivery Orchestration (PESCaDO)," 2010-2012, [Online; accessed 09-July-2012]. [Online]. Available: http://www.pescado-project.eu/

[28] L. Wanner, S. Vrochidis, S. Tonelli, J. Moßgraber, H. Bosch, A. Karppinen, M. Myllynen, M. Rospocher, N. Bouayad-Agha, U. Bügel, G. Casamayor, T. Ertl, I. Kompatsiaris, T. Koskentalo, S. Mille, A. Moumtzidou, E. Pianta, H. Saggion, L. Serafini, and V. Tarvainen, "Building an environmental information system for personalized content delivery," in *Proceedings of the 9th IFIP WG 5.11 International Symposium on Environmental Software Systems - Frameworks of eEnvironment (ISESS 2011), Brno, Czech Republic, June 27-29, 2011*, ser. IFIP Advances In Information And Communication Technology, vol. 359. Springer, 2011, pp. 169–176. [Online]. Available: http://link.springer.com/content/pdf/10.1007%2F978-3-642-22285-6_19

[29] M. Rospocher, S. Tonelli, L. Serafini, and E. Pianta, "Key-concept extraction for ontology engineering," in *Knowledge Engineering and Knowledge Management: 18th International Conference on Knowledge Engineering and Knowledge Management - EKAW 2012, Galway, Ireland, 2012*, vol. 7603. Springer Berlin Heidelberg, 2012, pp. 413–416. [Online]. Available: http://link.springer.com/content/pdf/10.1007%2F978-3-642-33876-2_36

[30] S. Tonelli, M. Rospocher, E. Pianta, and L. Serafini, "Boosting collaborative ontology building with key-concept extraction," in *Proceedings of 5th IEEE International Conference on Semantic Computing (September 18-21, 2011 - Palo Alto, USA)*, 2011.

[31] M. Rospocher, S. Tonelli, L. Serafini, and E. Pianta, "Corpus-based terminological evaluation of ontologies," *Applied Ontology*, vol. 7, no. 4, pp. 429–448, 2012. [Online]. Available: http://iospress.metapress.com/content/3356r67874680284/

[32] Organic.Edunet Consortium, "Organic.Edunet: A multilingual Web portal for sustainable agriculture & environmental education," 2007-2010, [Online; accessed 09-July-2012]. [Online]. Available: http://www.organic-edunet.eu/

[33] Organic.Lingua Consortium, "Organic.Lingua - Demonstrating the potential of a multilingual Web portal for Sustainable Agricultural & Environmental Education," 2011-2014, [Online; accessed 09-July-2012]. [Online]. Available: http://www.organic-lingua.eu/

[34] ProDe Consortium, "Progetto Dematerializzazione (ProDe)," 2010-2012, [Online; accessed 09-July-2012]. [Online]. Available: http://www.progettoprode.it/

[35] C. Di Francescomarino, M. Dragoni, M. Gerosa, C. Ghidini, M. Rospocher, and M. Trainotti, "Achieving interoperability through semantic technologies in the public administration," in *The Semantic Web: Research and Applications*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2012, vol. 7295, pp. 793–807.

[36] ProMo Consortium, "ProMo - Un Approccio Collaborativo e Agile per la Modellazione e il Monitoraggio di Processi di Business Basati su Servizi," 2012-2013, [Online; accessed 09-July-2012]. [Online]. Available: http://dkm.fbk.eu/index.php/ProMo

[37] C. Di Francescomarino, C. Ghidini, M. Rospocher, L. Serafini, and P. Tonella, "Reasoning on semantically annotated processes," in *Proceedings of the 6th International Conference on Service Oriented Computing (ICSOC'08)*, Sydney, Australia, 2008.

[38] OncoCure Consortium, "OncoCure - Information and Communication Technologies for Clinical Governance in Oncology," 2007-2009, [Online; accessed 09-July-2012]. [Online]. Available: http://www.cvast.tuwien.ac.at/projects/OncoCure/

[39] C. Eccher, A. Ferro, A. Seyfang, M. Rospocher, and S. Miksch, "Modeling clinical protocols using semantic MediaWiki: the case of the Oncocure project," in *ECAI workshop on Knowledge Management for Healthcare Processes (K4HelP)*, 2008.

[40] M. Rospocher, C. Eccher, C. Ghidini, R. Hasan, A. Seyfang, A. Ferro, and S. Miksch, "Collaborative encoding of asbru clinical protocols," in *Electronic Healthcare*, ser. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, M. Szomszor and P. Kostkova, Eds. Springer Berlin Heidelberg, 2012, vol. 69, pp. 135–143. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-23635-8_17

[41] eOnco Consortium, "eOnco Project," 2009-2011, [Online; accessed 09-July-2012]. [Online]. Available: https://dkm.fbk.eu/index.php/Projects

[42] J. Bao, L. Ding, D. L. McGuinness, and J. Hendler, "Towards social webtops using semantic wiki." in *International Semantic Web Conference (Posters & Demos)*, ser. CEUR Workshop Proceedings, C. Bizer and A. Joshi, Eds., vol. 401. CEUR-WS.org, 2008. [Online]. Available: http://dblp.uni-trier.de/db/conf/semweb/iswc2008p.html#BaoDMH08

[43] T. Kuhn, "AceWiki: A Natural and Expressive Semantic Wiki," in *Proceedings of Semantic Web User Interaction at CHI 2008: Exploring HCI Challenges*, 2008.

[44] S. Schaffert, "Ikewiki: A semantic wiki for collaborative knowledge management," in *1st Int. Ws. on Semantic Technologies in Collaborative Applications (STICA'06)*. [Online]. Available: http://www.wastl.net/download/paper/schaffert06_ikewiki.pdf

[45] F. Dengler, S. Lamparter, M. Hefke, and A. Abecker, "Collaborative process development using semantic mediawiki." in *5th Conference of Professional Knowledge Management. Solothurn, Switzerland, 2009.*

Move Better with tripzoom

Jing Bie[1], Marcel Bijlsma[2], Gregor Broll[3], Hu Cao[3], Anders Hjalmarsson[4], Frances Hodgson[5], Paul Holleis[3], Ynze van Houten[2], Koen Jacobs[6], Johan Koolwaaij[2], Diana Kusumastuti[1], and Marko Luther[3]

[1] University of Twente, Enschede, The Netherlands, {J.Bie, D.Kusumastuti}@utwente.nl
[2] Novay, Enschede, The Netherlands, {marcel.bijlsma, ynze.vanhouten, johan.koolwaaij}@novay.nl
[3] DOCOMO Euro-Labs, Munich, Germany, {broll, h.cao, holleis, luther}@docomolab-euro.com
[4] Victoria Institute, Gothenburg, Sweden, anders@viktoria.se
[5] Institute for Transport Studies, University of Leeds, F.C.Hodgson@its.leeds.ac.uk
[6] Locatienet, Amsterdam, The Netherlands, koen.jacobs@locatienet.nl

***Abstract*—The increase of urban traffic confronts individuals and transport authorities with new challenges regarding traffic management, personal mobility, sustainability, or economic efficiency. Existing resources cannot be arbitrarily extended without negative effects on pollution, costs, or quality of living and have to be used more efficiently. The SUNSET (Sustainable Social Network Services for Transport) project aims to improve this situation on a city-wide level by motivating users on a personal level to change their mobility behavior. To make personal mobility more sustainable, flexible, and rewarding for users, SUNSET combines mobility data and patterns from mobile sensing, a dynamic incentive system, and feedback from social networks. This paper describes how the tripzoom application implements this conceptual approach, discusses critical issues and outlines the forthcoming living lab evaluation in several European cities.**

Keywords–sustainable traffic; mobile mobility; mobile sensing; incentives; social networks; living labs

I. MOBILE MOBILITY

Urban environments provide many different challenges for transport and traffic authorities regarding sustainability (e.g., CO^2, air pollution), accessibility (e.g., congestion) or economic efficiency [1][2]. Existing traffic networks often operate at their limits and need to be extended or improved to meet the increasing mobility needs of an increasing number of people. Urban mobility management has to use existing resources more efficiently to improve personal mobility and to reduce congestion, accidents, and pollution at the same time. It also has to cope with conflicting interests. Road authorities want to optimize traffic to enhance accessibility and sustainability, while individual people want to move comfortably and save time and money.

Some approaches try to solve the challenges of urban traffic by extending the capacities of existing infrastructures and transport modalities. Other, more human-centered approaches target individual travelers and encourage them to change their travel behavior instead. They address personal goals like healthy living, flexibility, or sustainability, motivate travelers on a personal level to improve these goals and thus achieve system goals like reduced congestion, reduced air pollution or improved safety. Examples for novel mobility concepts that appeal to individual travelers and urban transport authorities alike are services for ride sharing (e.g., Avego [3]) car sharing (e.g., Whipcar [4], ZipCar [5] or GreenWheels [6]) or bike rental (e.g., DB Call-a-Bike [7], Velib [8] or Yokohama Bay Bike [9]) that show up in a growing number of cities all over the world.

Other solutions take advantage of mobile devices and applications to make travelling, transport and commuting smarter, more flexible and more beneficial for individual users. Mobile devices allow users to easily retrieve and communicate traffic-related information. They feature technologies like GPS or accelerometers to provide location data, suggest routes or detect user mobility. Mobile devices can also be highly personalized, context-aware [10] and persuasive [11]. They can influence the travel behavior of their owners by stimulating them in a personalized way and by providing incentives that match their motivations and preferences in the right place and at the right time.

The European FP7 project "Sustainable Social Network Services for Transport" (SUNSET) [12] is developing the tripzoom application [13] to investigate *mobile mobility* – solutions for personal mobility based on mobile devices and applications. With the tripzoom app (Fig. 1), SUNSET combines three means to improve personal mobility:

- **Personal Mobility Monitoring:** SUNSET uses mobile devices and their sensing capabilities to create individual mobility profiles with details about mobility patterns and transport modalities. These profiles allow users to zoom in on individual trips, frequent travels and special places and to re-visit them in detail.
- **Incentives:** tripzoom builds on this mobility data to provide users with incentives that match their interests to motivate the use of alternative means of transport and to encourage more sustainable travel behavior.
- **Social Networking:** Users can share mobility data and incentives with the tripzoom-community and existing social networks to encourage further improvements of their mobility behavior. For example, the mobile tripzoom app visualizes individual performances regarding costs or CO^2 footprint and shows users how well they perform compared to the community (Fig. 1).

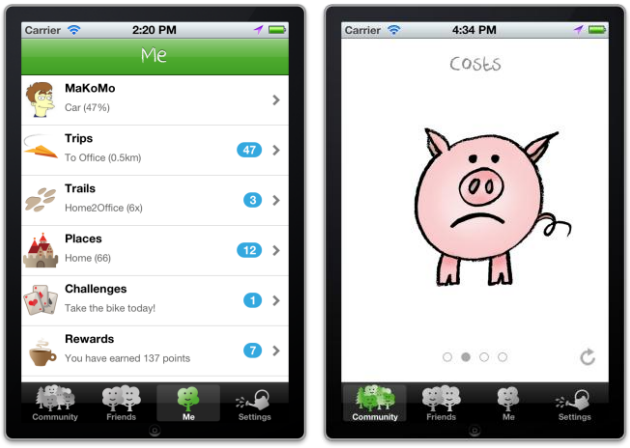


Figure 1. Mobile tripzoom app

The rest of this paper is organized as follows: After an overview of related work in Section II, Section III gives a general overview of the SUNSET approach and how it combines mobility monitoring, incentives and social networking to meet the challenges of modern traffic. Section IV describes the tripzoom application from a user's perspective, while Sections V and VI focus on technical details regarding system architecture and mobility sensing. Section VII concludes the paper with a discussion of our approach and an outlook to its evaluation in living labs in the cities of Enschede, Gothenburg and Leeds.

II. RELATED WORK

Our work on SUNSET and tripzoom is influenced by research on personal mobility, incentive mechanisms and living lab evaluations.

A. *Personal Mobility*

Most applications for personal mobility are related to information provisioning (e.g., Google Maps [14]) and mobility monitoring (e.g., [15][16][17][18]). Information can be provided to users via their mobile devices so that they can make better decisions regarding their trips. Such information involves trip planning, route guidance, and real-time traffic information (e.g., congestion, accidents).

Trip planning involves consulting maps, timetables, or online trip planners at home or relying on signs and fellow travelers on the road. As pointed out by Pitt et al. [19] the introduction of smartphones allows creating portable systems and multipurpose information appliances [20] that enable users to make sustainable decisions. With regard to personal mobility, smartphones make traffic related information, maps or timetables portable so that advanced tasks such as route planning can be done on the go. Mobile internet connections provide users with access to online trip planners and up-to-date traffic information, or allow them to arrange ride sharing [21] anywhere and at any time.

A common problem with most online trip planners is that they only accommodate single transport modalities. This makes the planner less useful for multi-modal trips, where the traveler has to consult several planners. Examples for multi-modal trip planners include the DB Navigator [22] or ReseRobot [23], that support multi-modal trip planning with plane, train, bus, metro, ferry, tram and walking. However, driving by car or bicycle is not accommodated, excluding P+R as an option.

In-car navigation devices are becoming a standard feature of most new car models. The on-board unit typically uses GPS to acquire position data and then locates the car using the unit's map database. Although stand-alone units are common, they are not popular for use in non-car modes, as carrying the device along can be a burden for the traveler. Route guidance systems for mobile phones, such as Google Maps [14], can minimize the carriage burden. The map database is normally stored on the server and accessed via the mobile phone's internet connection. Therefore, the guidance is always based on the latest database, significantly reducing inconveniences caused by outdated information.

Mobility monitoring is realized by sensing movement data and matching it with the traffic infrastructure. The sensing component tracks the position of users via GPS and cellular information, while the matching component fits these movement data into the context of the map database and establishes the trip-level characteristics, including origin, destination, timing or routing. Pattern recognition techniques are then utilized to identify the mode of travel and the role of users (e.g., driver, passenger), and to recognize places of interest over time by learning from the user's past behavior. Examples for personal mobility monitoring include TravelWatcher [24] and IYOUIT [25], which also supports the sharing of personal experiences on the go.

Personalization is often overlooked in existing mobile mobility systems. Although certain personal settings are allowed for filtering information, the majority of information is provided in the same way to all users. Guidance and advices are constructed as if they apply universally, disregarding the travelers' individual needs and preferences. The digital innovation contest TravelHack 2011 [26] tried to solve this problem by building prototypes for digital mobility services that present travel-related information according to personal user settings [27]. This indicates the potential for future mobility systems that allow users to adjust systems to their needs and preferences. The personalized approach in SUNSET is another attempt to bridge this gap in mobile mobility services.

B. *Incentives*

Over the past decade, there has been a growing interest in using incentives in transport both in commercial applications and government implementations to achieve sustainable policy aims. A number of EU-countries have implemented a range of different initiatives, like travel planning, public transport marketing, travel awareness campaigns or mobility management. Initiatives that share a common desire to change travel behavior through persuasion are collected under the term "soft measures", while "hard measures" involve enforcement and legislation.

The design of incentives in transport policies has a history of being more ambitious than the technology will

allow. For a long time there has been a vision and a desire to be able to offer personalized, multi-modal travel planning support as an incentive, but the technology has not been available to make it happen. Policy initiatives like Smarter Choices in the UK [28] have tried to make travelling less costly and more beneficial, sustainable or healthier for individuals using personalized travel planning or mobility management. However, these initiatives have often been relatively time-consuming and costly for local authorities and municipalities to implement.

Another, very simple and persuasive measure for motivating particular travel behavior are "monetized" inducements like discounted tickets for using public transport at certain times, price reductions for bulk or annual purchases of tickets, or discounted rates for entrance to tourist and visitor attractions for those who arrive by train or bus. The Spitsmijden initiative [29] in the Netherlands experimented with offering cash incentives to drivers who changed their car travel patterns and avoided travelling on a congested section of highway during the peak times. While this approach was effective and resulted in a change in congestion, it was also criticized for potentially introducing inequity into transport policy implementations [30].

Theory and practical evidence indicate that peer pressure can be a very influential incentive for changing travel behavior [31]. Many policy makers and commercial interests are experimenting with social networks like Facebook, Twitter or Foursquare, to encourage and support sustainable behavior. Examples include car sharing (e.g., [4][5][6]) or bike rental (e.g., [7][8][9]). Social interactions like sharing, collaboration or competition can be amplified by social networks. Tapping into these social interactions, applications can provide new forms of engagement and stronger incentives.

Within the SUNSET project, a number of key approaches to incentive design have been explored [32]. We have found out that social network based sharing of incentives is attractive to some demographic groups but not all, that point-based incentives have to be exchangeable to generate loyalty, and that feedback is a mechanism that can be mainstreamed into everyday travel behavior. tripzoom offers the possibility of bringing together a number of novel and inventive concepts in incentive design through the use of dynamic, personalized, mobile social media to develop sustainable travel behavior.

C. *Living Lab Evaluation*

tripzoom will be evaluated in the real-life context of living labs [33] in Enschede, Gothenburg and Leeds. Eriksson et al. [34] define living labs as a research and development methodology whereby application enhancements, services, or products are created and validated in collaborative, multi-contextual empirical real-world settings. All stakeholders in a product, service or application participate directly in the development process. For SUNSET, this includes end-users and municipalities. Living lab experimentation strives for the same level of observation as is common in, for example, a usability lab, but in an organic, multi-contextual space. This means that customers participating in a living lab are observed across many aspects of their lives, such as their roles as citizens, workers, at home or travelling. As such, living lab studies provide user feedback of high ecological validity. In SUNSET, the living labs will also collect high-value and real-life mobility data about their mobility behavior and patterns.

III. THE SUNSET APPROACH

The European FP7 research project SUNSET [12] takes a new approach to urban mobility management and personal mobility. For that purpose, its consortium contains partners from the complete value chain: providers of location-based services, mobile operators, local authorities, green mobility providers and research centers. The focus of SUNSET lies on urban mobility with a fine-grained maze of roads (as opposed to long-distance highways) and on commuters with good knowledge of their environment, but with limited overview of dynamic situations and routine behavior.

The SUNSET approach builds on mobile devices, personal mobility monitoring, a dynamic incentive system and information sharing over social networks to optimize mobility for individual users and to encourage the adoption of more flexible, sustainable and rewarding ways of traveling. This approach includes research on the recognition of individual mobility-patterns and transport modalities, the effects of different incentives to encourage users to change their travel behavior as well as mechanisms for building social communities and sharing travel-related information.

The SUNSET approach envisions an eco-system, in which individual travelers, communities, city authorities and 3rd party service providers interact with each other to optimize personal mobility (Fig. 2). Individual travelers provide the SUNSET eco-system with information about their preferences as well as mobility data about travels, transport modalities or places. In return, travelers can take advantage of personalized recommendations and services to optimize their personal mobility, share travel related information on social networks and receive incentives for improved travel behavior.

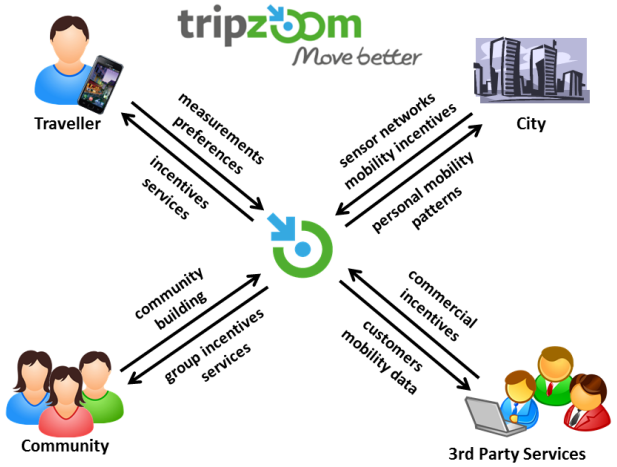


Figure 2. Stakeholders in the SUNSET eco-system

City authorities receive detailed, personal mobility data from travelers that are relevant for the assessment of current infrastructure use and future mobility needs. To optimize this data, cities can offer incentives to travelers. The information is targeted at individual travel behavior, and thus allows road authorities to fine-tune their transport policies and individual travelers to meet their personal objectives.

Third party service providers can tap into the wealth of mobility data to create novel services and offerings for travelers. They can also integrate with other parties and stakeholders through common incentive structures.

Communities, like participants of a car sharing system or employees of a company, can receive special group incentives and improve their networking.

A. Mobility Monitoring

To optimize personal mobility, individuals or groups need to be identified that are relevant for the optimization targets under consideration. Therefore, personal mobility sensing that uses mobile phones to track how travelers move, plays a fundamental role in the SUNSET system to gain profound insights into mobility behavior. One example is the detection of travel modalities, which combines orientation sensors, location data, and road network information to estimate the mode of transport in a reliable way. Other than systems that build on infrastructure data alone, e.g., on road sensors like NDW [35] in the Netherlands, SUNSET is able to target individual travelers to initiate a change in behavior in accordance with a global optimization of travel behavior across a whole city. For that purpose, the mobility behavior of the entire community or specific groups can be analyzed by aggregating mobility data across the respective users.

B. Incentives

One approach SUNSET employs to trigger behavior change is to provide incentives for specific target groups in specific mobility situations. Incentives may stem from all parties of the SUNSET ecosystem and may contain information about the current and future status of the transport infrastructure or about travel alternatives. To find the most effective types of incentives, we conducted an analysis of individual travel behavior [36]. This research implies that influential incentives should be based on the following aspects.

- **Time:** Travelers can save time, use it more efficiently, control it or plan trips in a better way.
- **Money:** Travelers can save money, for example with coupons or discounts on transportation tickets.
- **Information:** Travelers can receive useful (real-time, personalized) travel-related information about progress, accidents, alternative routes, etc.
- **(Social) Recognition:** Travelers can show (off) how green and healthy they are and give or receive feedback from other members of the community.

In SUNSET, incentives appear as challenges defined by a set of rules that users have to fulfill to earn a reward (e.g., take the bike 3 times a week instead of the car to get a discount on bike maintenance financed by a bicycle store). SUNSET promotes an abstract point scheme to implement rewards and to facilitate the integration of 3rd party incentive offers. This includes the possibility to exchange point for other types of rewards, for example monetary discounts. Furthermore, points also enable users to track their progress and to compare it with friends and colleagues.

To evaluate the effects of different (combinations of) incentives, SUNSET foresees dynamic and controllable incentive management. It allows for generating and placing incentives in real-time during the entire runtime of the system. This means that weak impact or unintended side effects of incentive offers can be detected early and incentive operators, such as city representatives, can alter incentives and the corresponding awards accordingly. For example, city operators can quickly set up incentives to prevent traffic jams after a large concert by motivating visitors to use public transport. In case of temporary overloading, operators can quickly react and issue another incentive to motivate people to stay in the area longer (by offering a free coffee nearby).

C. Social Networking

SUNSET also uses existing social networks, like Facebook, Foursquare, or Twitter, to motivate users to change their travel behavior. Their social sharing infrastructure is ready for users to share and advertise their rewards, achievements, and progress with their friends and other people and get feedback from them. Social networks can also support competition as a gamification feature [37] and create a more playful experience. They can be powerful tools for inducing behavior change as users can compete with each other by sharing and comparing their achievements. Social networks can also trigger conformance to group behavior, especially within a community of people with similar goals and interests.

IV. TRIPZOOM

The tripzoom system is a concrete implementation of the general SUNSET approach and consists of three parts: a mobile application (app) that encourages users to improve their mobility behavior, a Web portal that introduces tripzoom to novel users, and a city dashboard that allows providers to manage tripzoom and its different features.

A. Mobile App

The tripzoom app for iPhones and Android devices is the center of the tripzoom system and makes its main features available to users. To provide these features, the app uses the sensing capabilities of modern mobile devices (e.g., GPS, accelerometer) to detect, measure and track the movements of their users. This raw sensor data is filtered, refined, and finally turned into useful mobility data about places, trips, or transport modalities. This data can later be used to derive the travel behavior of users and give recommendations for how to improve it. The app also visualizes mobility data in different ways and allows users to interact with them. The different features for visualization and interaction are represented by the tabs “Community”, “Friends” and “Me” in the user interface of the app (Figs. 1 and 3).

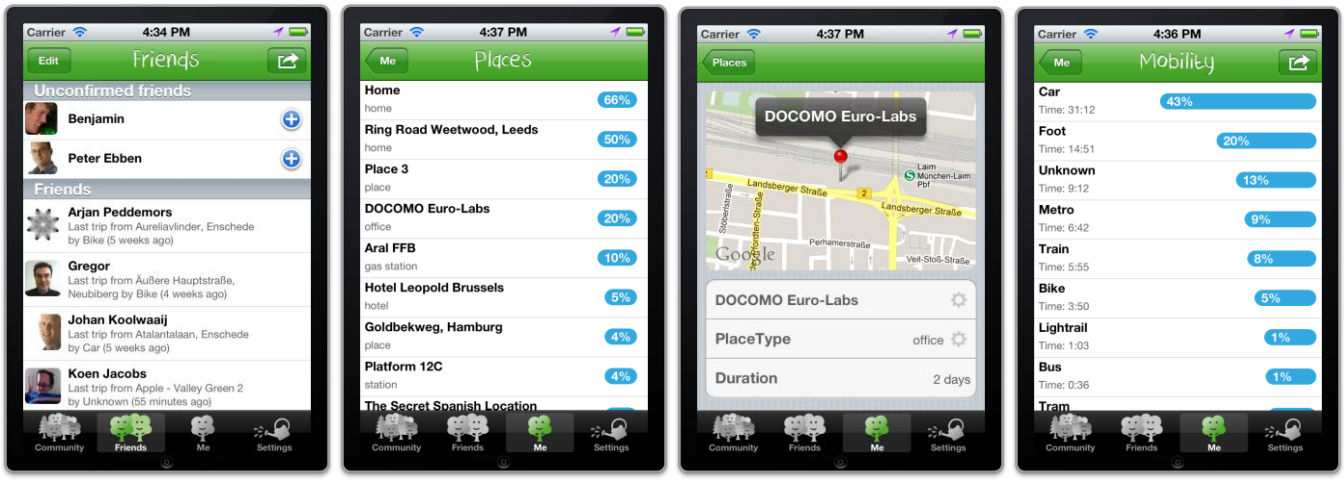


Figure 3. Visualizing mobility data in the tripzoom app

- **Community:** This tab gives an overview of how well the user of the app performs compared to the whole tripzoom community (Fig. 1, right). Four different visualizations illustrate this performance regarding saved money and CO^2, health and collected points (Fig. 4). If a user performs better than the community average, the respective illustration is more positive, if he performs worse, it is more negative.
- **Friends:** This tab shows the list of friends in the tripzoom community. The list can be managed on both the mobile app and in the Web portal. Every list entry provides additional information about a specific friend, for example about his last trip or trip statistics.
- **Me:** This tab contains all information that is related to the user of the tripzoom app (Fig. 1, left). It provides detailed information about his mobility profile, including visited places, individual and frequent trips ("trails") or statistics on travel modalities (Fig. 3). It also manages challenges that a user tries to achieve and gives an overview of rewards he has earned.
- **Settings:** This tab contains different options for the management of the tripzoom app, including settings for privacy, profile information, or sharing data.

B. *Web Portal*

The tripzoom Web portal is the main entry point into the tripzoom experience for new users and consists of two main parts: The landing page advertises tripzoom to interested users, explains its features, provides background information and links to the App Store and Google Play from where users can download the tripzoom app for iPhones and Android devices. The landing page is also a first step into the tripzoom community as it provides links to tripzoom pages on social networks (e.g., Facebook, Twitter, Google+ and Foursquare) and shows updates from Facebook and Twitter feeds. In addition, a community feature provides anonymous and aggregated real-time statistics about the performance of users in the three living lab cities regarding the number of trips or the amount of CO^2 that is saved through improved personal mobility.

On the landing page, users can log in to their accounts or create new ones. The portal provides all functionalities to let users maintain their accounts, edit their profiles, or manage settings. As the portal puts a focus on social features, users can create blogs, send messages, invite friends and manage their relationships with them, e.g., by categorizing them as "family members", "close friends" or "colleagues" (Fig. 5). An activity stream gives a dynamic overview of interesting actions that users carry out in the portal, such as creating new blog posts or accepting friend requests.

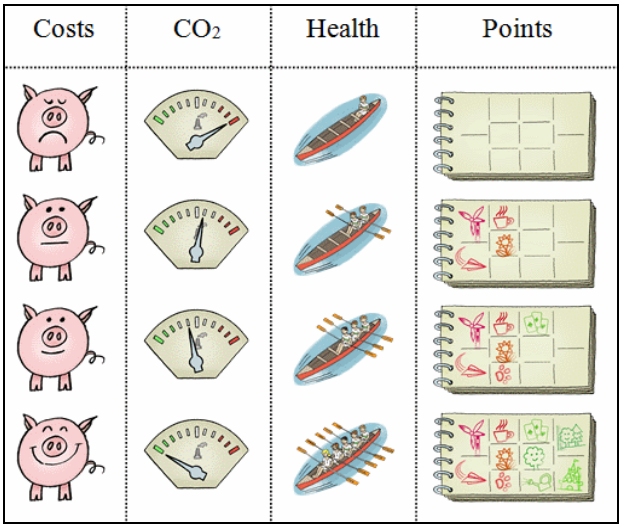


Figure 4. Visualizations of user performance

C. *City Dashboard*

Next to the Web portal and the mobile app, tripzoom offers a service called city dashboard that allows city authorities to monitor and manage the incentives and reward system. They can add new incentives and specify for which

target group and in which situations these will be triggered. Furthermore, they can get an overview of the current and past mobility situation in the city to design suitable and effective incentives. The city dashboard offers a live view on various aggregated statistics and anonymized data of the users within a city with which operators can check and evaluate the current state and impact of incentive measures. This approach can be used to influence overall mobility behavior (e.g., fewer cars in the city) or to target specific goals, e.g., optimize the use of alternative travel modalities before and after large events.

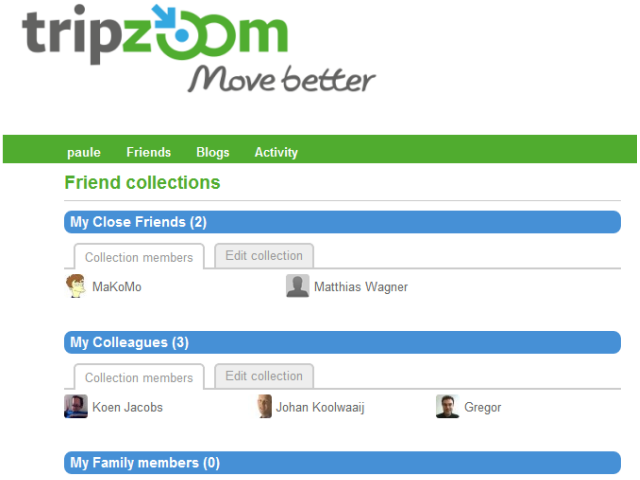


Figure 5. Friend management in the tripzoom Web portal

V. ARCHITECTURE AND IMPLEMENTATION

tripzoom follows a client-server architecture offering a service infrastructure that provides a set of core mobility and social networking services guarded by a security layer. As described above, users interact with one of the smartphone clients for iPhone or Android or with a Web user interface. City controllers can make use of a city dashboard view, which provides access to city wide mobility information and incentive control. Finally, 3rd party services such as the social networking applications Facebook and Twitter expand the reach of tripzoom information and allow for personalized sharing of mobility information.

In the following, we provide an overview of the general architecture with its main components and some detail about the central functionality, the collection and interpretation of the users' movement data to build mobility profiles. The tripzoom server side consists of a set of core services providers, Web user interfaces, and programming interfaces to 3rd party services, data, and applications (Fig. 6). It delivers core services including incentives, mobility monitoring, and basic social networking services.

- **Personal Mobility Store (PMS):** Collects raw measurements from mobile clients and preprocesses them as input to algorithms for mobile pattern detection.
- **Mobility Pattern Detector (MPD):** Receives data from the PMS and employs sophisticated algorithms to detect patterns for individuals, groups, places, regions, routes, or vehicles such as bus lines or taxis.
- **Relation, Identity, and Privacy Manager (RIP):** Provides a homebuilt social network implementation and organizes the privacy policies of users based on their social relations or ad-hoc groupings computed by the MPD.
- **Social Network Connector (SNC):** Connects the internal social network (RIP) with existing social networks such as Facebook or Twitter to facilitate user registration, information sharing (e.g., a notification on the successful completion of an incentive), importing contacts, or showing visualizations from the MPD.
- **Incentives Market Place (IMP):** Provides a platform to offer rewards, recognition or real-time feedback as incentives to encourage travelers to improve their travel behavior with respect to the system's and an individual's objectives. The IMP matches available challenges with mobility patterns from the MPD, individual user profiles and preferences from the RIP, general transport information, and can publish performance and events using the SNC.
- **Context Harvester (CH)**: Harvests all information required to populate the user's buddy list from all server-side components, such as the RIP for the user profile, the PMS for the last location and trips, the MPD for mobility patterns and the IMP for rewards gained with incentives.

Additionally, the tripzoom server provides a Web portal where mobile users can view their mobility profiles online. 3rd party applications can request access to services provided by the core service provider components after obtaining the user's consent. The distributed system is implemented building on Representational State Transfer (REST) and JavaScript Object Notation (JSON). This supports loose-coupling between components, clear application programming interfaces, and independence of platform and programming language. The security layer uses OAuth, a simple mechanism to publish and interact with data that needs access control. OAuth is a de facto standard used by many systems such as Facebook or Twitter. The tripzoom social network features such as the Social Network Connector build on the open source social networking engine ELGG [38]. Using such a platform has the advantage that users are not required to use a specific network and it simplifies the integration with other functionalities that the portal and the mobile application offer.

The mobile devices of tripzoom users serve as clients to provide mobility data and to present transport information and incentives to guide the mobile user's travel behavior. Correspondingly, the mobile client comprises components for mobile sensing, graphical user interfaces and secure communication with the server (Fig. 6). The sensing component is responsible for gathering mobility sensing data from sensors, to preprocess the data thereby removing redundant and low quality items, and then to upload them to the tripzoom server. Some sensing data elements such as location and battery level can be obtained from onboard

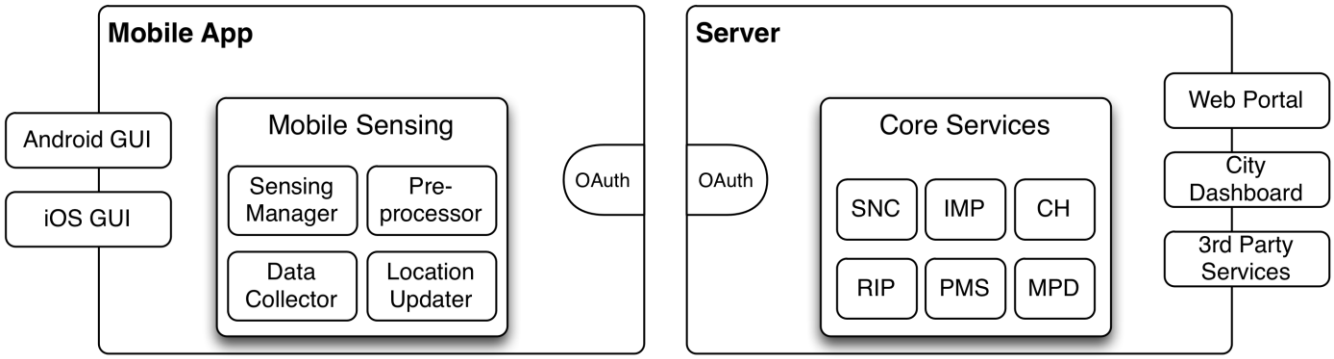


Figure 6. Architecture of the tripzoom system

mobile device sensors. Other data can be incorporated from external sensors, such as step counters or bike sensors. Currently, the tripzoom app supports a range of sensors based on the ANT+ protocol such as the Fitbit Tracker [39] or the Wahoo Cycling Speed/Cadence Sensor [40]. There are four sub-components in the mobile sensing pipeline.

- **Data Collector:** Gathers raw measurements from built-in mobile device sensors as well as externally connected sensors.
- **Sensing Preprocessing:** Applies algorithms to reduce the noise and size of the gathered data, to recognize stationary and travelling situations, and computes initial trip modalities based on the gathered data and the type of the corresponding sensor source.
- **Location Updater:** Decides when to upload the sensing data to tripzoom servers and uploads them when appropriate.
- **Sensing Manager:** Activates, deactivates and adjusts the available sensors and sampling rates to optimize energy consumption as well as data quality.

The social and UI components provide the user interface showing the users' mobility profile (mobility profile visualization), live status of their friends (buddy status and social networking), and receive incentives for a healthier and more cooperative mobility behavior (incentive presenter). A security and communication component is responsible for authenticating mobile users to access their sensitive personal mobility data and to synchronize the mobility, incentive, and social network data between tripzoom mobile clients and the tripzoom server. Upload and download links to push data to and query data from the tripzoom server is implemented by a communication manager module.

Another important component to keep the users up-to-date and to be able to request their feedback on the spot is integrated in the mobile notifications component. This is used to published new incentives, to inform about earned rewards, and to send experience sampling questions. As it is based on the respective mobile operating system's notification mechanism, it can deliver messages even when the phone is in sleep mode.

VI. Mobility Sensing and Mining

tripzoom uses various technologies in mobile phones (e.g., GPS, Wi-Fi, accelerometer) to detect or deduce the 4 Ws of personal mobility: when do people move, where (via which route), with whom, and using which modality (e.g., bike, car). While personal mobility data allows detailed, individual tracking, only a small part of the population can be expected to participate. Still, the goal is to get a comprehensive mobility profile per traveler (Fig. 3) covering a 24/7 period. The use of mobile phone sensors and battery power has to be carefully balanced with the measurement accuracy to be achieved. In the following, we provide more details about how tripzoom senses, gathers, and analyses data from mobile phone sensors keeping the additional energy consumption to a minimum. Fig. 7 gives an overview of the three main processing stages for mobility sensing and mining:

- **Mobile Sensing:** Location data are the key to construct a person's mobility profile. Various data sources from a smartphone's sensors need to be gathered, outliers must be removed, and data transmission optimized.
- **Trip Reconstruction:** This continuous but often noisy location data has to be categorized as a series of static periods and trips with potentially different travel modalities. To get the actual route of the trip, the location trace is mapped onto the road or rail network.
- **Pattern Mining:** Location points and traces are examined to identify frequently visited places and regular trails, such as commuting routes. This information is crucial for isolating trips that will have the largest impact on the user's mobility behavior when changed.

A. *Mobile Sensing*

Information from mobile device sensors (e.g., GPS, Wi-Fi, and GSM measurements) is combined into one location model that can deal with the varying accuracy of location measurements. These location signals contain a lot of noise and outliers. Especially, the accuracy of GSM data is usually low and the reported locations may be hundreds of meters away from the actual locations of the phone.

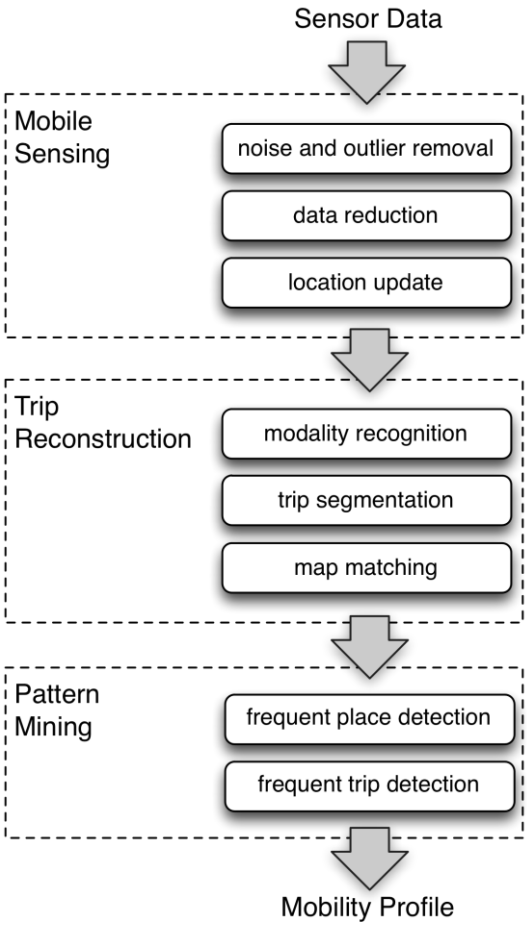


Figure 7. Mobility sensing overview

We designed a location update approach that combines location updates to the tripzoom server with noise removal and data reduction. For a smartphone with GPS, Wi-Fi, and GSM location sensors (in case of Wi-Fi and GSM, locations are computed by a central service provider like Google or Skyhook [41] based on the sensing data provided by the smartphone), tripzoom registers a location change listener to continuously receive location data.

As shown in Fig. 8, our approach considers four scenarios for a location update. If, as shown in Fig. 8a, the new location *q* is far away from the old location *p* and the following location *r*, but the time difference is little (i.e. the calculated speed from *p* to *r* through *q* is high), the update *q* is considered an outlier. Subject to further empirical studies, we use 50m/s as the threshold of bad data selection. Fig. 8b depicts a common scenario that the new location *q*'s error range is much worse than that of *p*. A common example is that *p* is given by Wi-Fi or GPS, while *q* is a GSM location. Since the error ranges of *p* and *q* overlap, *q* will be discarded as a low accuracy and unnecessary location update. The missed location will be compensated by the map matching approach detailed in the next section. If the new location *q* is more accurate (as is depicted in Fig. 8c) or *q*'s error range does not overlap with the previous location *p* (as in Fig. 8d), the new location *q* is accepted as a good location update and sent to the tripzoom server for further processing.

To keep the amount of data to be processed and transferred small (thus reducing energy consumption, storage space, and transfer times), location updates within a certain threshold distance are also discarded as these add little information for the subsequent route reconstruction.

To implement the trip detection in a battery-efficient way, the tripzoom app listens to 'significant' location changes using Wi-Fi (preferred) or GSM location updates and only enables the battery-hungry GPS if it has reasons to assume that the user has left the (static) location. Attempts to acquire a GPS fix are kept short. If there is no fix within that time, the GPS is turned off and will be re-enabled only after some time. By varying those time intervals, a trade-off is made between consuming battery power and attempting to record the trip as detailed as possible. Our initial tests within a limited user community have shown that battery life can be brought to 20 hours on average, with quite some variations related to differences in coverage (GSM, GPS, etc.) and mobile phone use.

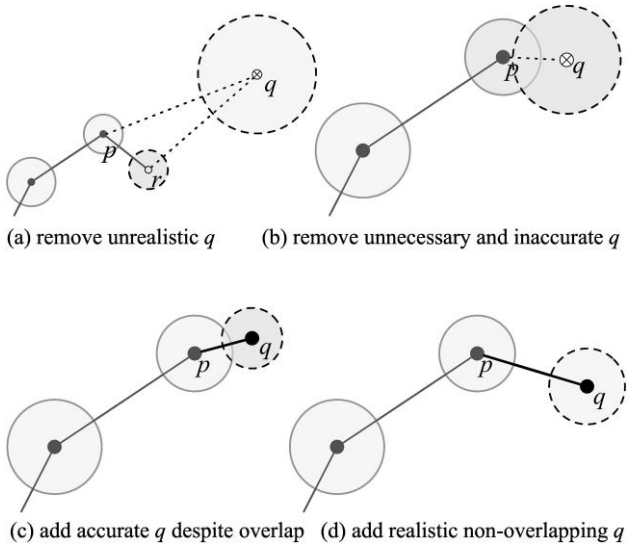


Figure 8. Location update

B. *Trip Reconstruction*

The task of trip reconstruction is to estimate the route traveled, deduce the places of origin, destination and intermediate stops, as well as to infer the transport modalities used per trip segment, based on the recorded data. Therefore, on the back-end, the goal is to further enhance the data, to apply smoothing and outlier detection to the location data, and to improve trip timings. To gather information on overall modality use and reward calculation, trips need to be stitched and split such that each segment covers a single modality only. To facilitate modality detection, each route segment is mapped onto the underlying infrastructure network. Our algorithms estimate the modalities of each trip based on the location measurements as well as derived parameters such as

speed pattern, infrastructure usage, and origin and destination of the trip. Trip modality recognition applies rules to all gathered information and calculates modality likelihoods. For example, a trip going from one airport to another with an average speed above 150km/h will likely have been made by plane. Users travelling on roads closed for car traffic are likely to walk or cycle depending on their speed pattern, etc. Still, it is important to allow users to manually adapt the recognized modality, in case of any error. The current system has an accuracy of about 70%, which already saves considerable effort compared to manually labeling all trips and is similar to other state-of-the-art approaches [42]. Its algorithm works better on "walking", "cycling", "car", "plane", and "train" than on other modalities including "bus" and "tram" as these share various properties with similar modalities such as "car" or "train". An improved approach for recognizing these modalities has been designed, which takes into account bus and tram stops as well as train stations. Trips passing along these places with significant differences between both trip segments, e.g., in speed, will be automatically split at that location, thereby improving the modality detection.

Both trip reconstruction and modality detection strongly rely on map matching. Map matching is the process where the geographic trace in terms of latitude, longitude, and accuracy of the location measurements of a trip is mapped onto the infrastructure network in terms of streets, paths, rails, ferry, and air connections. We have made our own implementation based on existing open-source components [43], and each trip submitted to the PMS is map matched, resulting in a reliable estimate of which infrastructure was actually used to make this trip. This infrastructure usage is then input for the modality detection in the sense that a higher percentage of bus lane usage will favor the "bus" modality, and higher percentage of paths only suited for walking and cycling will decrease the likelihoods for all four-wheeled modalities. Discontinuities in infrastructure usage, like transitions between road and rails, or between rails and air, are proper triggers for the trip splitting algorithm. And the higher the quality of the information on the lowest trip level is, the higher will be the quality of every piece of (pattern, profile or community) information derived from all trips made in SUNSET's living labs.

C. Pattern mining

To challenge travelers into new – and more optimal – behavior, it is very important for the tripzoom system to know what someone's current mobility behavior is. Based on this knowledge, tripzoom can notice deviations from the normal routine, which might be the traveler's response to an attractive incentive. This way, the effectiveness of tripzoom's incentives can be measured. Personal places, regular trips, called trails in the following, and modality statistics are key components of personal mobility profiles.

Personal places are geographical places that a traveler either visits frequently or for a longer period of time. These personal places will be automatically linked to nearby public places, and the traveler can also adapt naming and typification of these places manually. These places are then used to assign origin and destination to the personal trips, thereby adding semantics to these trips, as it will not only be classified as a trip from street "A" to street "B", but as a trip, e.g., from "home" to "work" with the objective to "go to work", or to a "shopping mall" with the objective to "go shopping". Personal places are automatically detected by analyzing the static periods of a traveler between two trips. Smart smoothing is applied, and the spikes in the 2D geographical duration-of-stay-histogram are then recognized as personal places in terms of (latitude, longitude, radius). These are used to either update information on the previously detected places, or to add newly detected places. Per place, numerous statistics are computed and persisted, including absolute and relative duration of stay, and their distributions over the hours of the day and the days of the week (Fig. 9).

Figure 9. Personal place statistics

Once tripzoom has knowledge about the personal places of a user, it can also start deducing trails. This is done by clustering all detected trips with the same origin, destination, route, and modality. The clustering is not done on the geographical level, but on the level of Open Street Map [44] infra segments, cancelling out short detours and potentially missing detailed information in the beginning of the trip. All trails with the same origin and destination provide a good overview over all routes and modalities a traveler uses to get from that origin to that destination, and when and how often exactly. Furthermore, all newly detected trips are – if possible – assigned to one of the existing trails consumption if a traveler is on his way to office with a known modality and route for the nth time.

The *personal modality statistics* provide an overview of how much time, distance, costs, and emissions a traveler spends per modality over different periods of time. This is done by analyzing all trips in that period of time and computing the total amounts per modality and other conditions such as whether it is rush hour or not, whether it is raining or not, etc. A perfect capture of a traveler's normal mobility behavior can be used to add someone to the target group for specific incentives. With the modality statistic at hand, the appropriate triggers can be provided automatically by the tripzoom system.

VII. Conclusion and Future Work

The tripzoom app supports a personalized mobility monitoring and dynamic incentive system within a social network context. The major contribution of tripzoom lies in the combination of the following aspects:

- tripzoom implements a mobility sensing and profiling service that is accurate and battery friendly. This allows users to monitor their individual mobility behavior 24/7.
- tripzoom supports the operation of innovative incentive programs that will encourage sustainable mobility. The sustainability goals can be achieved through the adaptations in traveller behaviour regarding destination, timing, mode, and route choices.
- tripzoom provides a social networking environment that is attractive to the e-generation. It can be utilized to promote identity recognition and synergy of travellers on sustainable travel behaviour.

The tripzoom system has the potential to meet mobility challenges our society is facing nowadays, with an ever increasing emphasis on efficiency and sustainability. The on-going living lab experiments will not only validate the operational maturity of the app, but also shed light on the impact of tripzoom on the policy goals commonly used in Europe and beyond, such as congestion and emission. The living lab experiments [45] take place in three European cities, where tripzoom is put to the test in the real world: Enschede, the Netherlands; Gothenburg, Sweden; and Leeds, U.K. Participants are recruited from these cities to be active users of tripzoom. Mobility experts at the municipal governments act as living lab coordinators and determine the issuance of incentives among the user groups.

The primary objective of the living labs is to evaluate the effectiveness of tripzoom and its incentive system and how much it can influence the mobility behavior of its users. The evaluation is based on monitoring individual changes by particular population subgroups in particular contexts. The potential impact of these changes at system level can then be estimated. These results will provide valuable input for the improvement of existing incentives and the identification of new ones. Besides the evaluation of incentives, the living labs will provide user feedback of high ecological validity for the operational success of tripzoom and assess the impact of tripzoom on local traffic policies. In the end, the living labs will establish a continuously growing and developing initiative in the cities involved, where a community of citizens, companies and municipalities is formed for realizing their objectives.

Various issues have arisen during the development of tripzoom. On-going research within the SUNSET consortium strives for the further enhancements of the following issues.

- **Robust Mobility Detection vs. Battery Efficiency:** Trip sensing requires the detection of movements and the subsequent matching into mobility patterns. Firstly, the app may fail to detect the mobile phone's movement. Secondly, given that some movements are registered, the app may not fully recognize the trip's origin, destination, timing, and mode, or recognize them in the wrong way. There is a trade-off between detection accuracy and battery consumption. Our research will strive for a highly accurate detection system that consumes as little battery power as possible.
- **Effectiveness of Incentives:** Incentives provide stimuli to individual travelers for changing their behavior. Behavioural adaptation is dependent on the types and operational characteristics of these incentives. This relationship can be used to derive the most effective incentives, which are expected to vary depending on the traffic situation and the group of travellers.
- **User Friendliness:** Usability of an app strongly affects its attractiveness to new users as well as the retention of existing users. A user friendly app needs to be effective, efficient and satisfactory in its specified context of use. For tripzoom, user interface design, user documentation, and error prevention will be further enhanced to improve the user experience.
- **Attractiveness:** Participating in a living lab should be attractive to the average commuter and citizen. It will be interesting to know the share of people who want to contribute to a 24/7 living lab for the common good, especially in comparison to the share that is required to draw valid conclusions on city mobility as a whole.

REFERENCES

1. P. Holleis, M. Luther, G. Broll, H. Cao, J. Koolwaaij, A. Peddemors, P. Ebben, M. Wibbels, K. Jacobs, and S. Raaphorst, "tripzoom: A System to Motivate Sustainable Urban Mobility," Proc. International Conference on Smart Systems, Devices and Technologies (SMART '12), ThinkMind, May 2012.
2. M. Naphade, G. Banavar, C. Harrison, J. Paraszczak, and R. Morris, "Smarter cities and their innovation challenges," Computer, vol.44, no.6, pp.32-39, June 2011.
3. Avego website. www.avego.com. Accessed: 20.12.2012
4. WhipCar website. www.whipcar.com. Accessed: 20.12.2012
5. Zipcar website. www.zipcar.com. Accessed: 20.12.2012
6. Greenwheels website. www.greenwheels.com. Accessed: 20.12.2012
7. Call a Bike website. www.callabike-interaktiv.de. Accessed: 20.12.2012
8. Velib website. en.velib.paris.fr. Accessed: 20.12.2012
9. Yokohama Bay Bike. http://docomo-cycle.jp/yokohama/top. Accessed: 20.12.2012
10. G. D. Abowd, A. K. Dey, P. J. Brown, N. Davies, M. Smith, and P. Steggles, "Towards a Better Understanding of Context and Context-Awareness," Proc. of International Symposium on Handheld and Ubiquitous Computing (HUC '99), Springer-Verlag, Sep. 1999, pp. 304–307.
11. B. J. Fogg and D. Eckles (eds.), "Mobile Persuasion: 20 Perspectives of the Future of Behavior Change," Stanford Captology Media.
12. SUNSET project website. www.sunset-project.eu. Accessed: 20.12.2012
13. tripzoom website. www.tripzoom.eu. Accessed: 20.12.2012
14. Google Maps. maps.google.com. Accessed: 20.12.2012
15. J. Ryder, B. Longstaff, S. Reddy, and D. Estrin, "Ambulation: A tool for monitoring mobility patterns over time using mobile phones," Proc. of the 12th IEEE Int. Conf. on Computational Science and Engineering (CSE '09), IEEE Computer Society, 2009, pp. 927–931.
16. S. Wang, C. Chen, and J. Ma, „Accelerometer based transportation mode recognition on mobile phones," Proc. of the 2010 Asia-Pacific Conference on Wearable Computing

Systems (APWCS'10), IEEE Computer Society Washington, DC, USA, 2010, pp. 44–46.

17. P. Siirtola and J. Röning, "User-independent human activity recognition using a mobile phone: Offline recognition vs. real-time on device recognition," Advances in Intelligent and Soft Computing, volume 151, pages 617-627.
18. J. J. Guiry, P. Van De Ven, and J. Nelson, "Orientation independent human mobility monitoring with an android smartphone," Proc. of the IASTED International Conference on Assistive Technologies (AT'12), ACTA Press, 2012, pp. 800–808.
19. L. F. Pitt, M. Parent, I. Junglas, A. Chan, and S. Spyropoulou, „Integrating the smartphone into a sound environmental information systems strategy: Principles, practices and a research agenda," Journal of Strategic Information Systems, volume 20, 2011, pp. 27–37.
20. S-J. Hong and K. Y. Tam. "Understanding the Adoption of Multipurpose Information Appliances: The Case of Mobile Data Services," Information Systems Research, volume 17, number 2, 2006, pp. 162–179.
21. N. D. Chan and S. A. Shaheen, "Ridesharing in North America: Past, Present, and Future," The Transportation Research Board Annual Meeting 2011. Transport Reviews, volume 32, number 32, 2012, pp. 93–112.
22. DB Navigator website. www.bahn.de/iphone. Accessed: 20.12.2012
23. Reserobot website. www.reseplanerare.resrobot.se. Accessed: 20.12.2012
24. W. B. Teeuw, J. Koolwaaij, and A. Peddemors, "User behaviour captured by mobile phones," Proc. of the Workshop on Interactive Human Behavior Analysis in Open or Public Spaces (INTERHUB'11), Springer Berlin, 2012, pp.81-90.
25. S. Böhm, J. Koolwaaij, M. Luther, B. Souville, M. Wagner, and M. Wibbels, "Introducing IYOUIT," In Proc. of the Int. Semantic Web Conference (ISWC'08), Springer, 2008, pp. 804-817.
26. TravelHack website. www.travelhack.se. Accessed: 20.12.2012
27. A. Hjalmarsson and D. Rudmark, "Designing Digital Innovation Contests," In K. Peffers, M. Rothenberger, B. Kuechler (Eds.) Proc. of the 7th Int. Conf. on Design Science Research in Information Systems: advances in theory and practice (DESRIST'12). LNCS volume 7286, pages 9–27. Springer Berlin, 2012.
28. Smarter Choices – Department for Transport website. www.dft.gov.uk/topics/sustainable/smarter-choices. Accessed: 20.12.2012
29. Spitsmijden website. www.spitsmijden.nl. Accessed: 20.12.2012
30. M. C. J. Bliemer and D. H. van Amelsfort, "Rewarding instead of charging road users: a model case study investigating effects on traffic conditions," European Transport, number 44, 2010, pp. 23–40.
31. A. Binsted and R. Hutchins, "The role of social networking sites in changing travel behaviours," TRL Published Project Report, number PPP599, TRL, 2012.
32. D. Kusumastuti (editor), "Impact of Incentives," EU Project SUNSET, Deliverable D3.3. 2012.
33. EC-ISM European Commission – Information Society and Media. Living Labs for User-driven Open Innovation – An Overview of the Living Labs Methodology, Activities and Achievements. Luxembourg: Office for Official Publications of the European Communities, 2009. http://ec.europa.eu/informationsociety/activities/livinglabs/docs/brochure_jan09_en.pdf. Accessed: 20.12.2012.
34. M. Eriksson, V.-P. Niitamo, and S. Kulkki, „State-of-the-art in utilizing Living Labs approach to user-centric ICT innovation - a European approach," www.vinnova.se/upload/dokument/Verksamhet/TITA/StateoftheartLivingLabs Eriksson2005.pdf. Accessed 20.12.2012.
35. Nationale Databank Wegverkeersgegevens website. www.ndw.nu/en. 20.12.2012
36. A. Ma (editor), "Feasible and Potentially Successful Incentives," EU Project SUNSET, Deliverable D3.4, 2012.
37. G. Zichermann, "Gamification by Design," O'Reilly, 2011.
38. Elgg – Open Source Social Networking Engine website. http://elgg.org. 20.12.2012
39. Fitbit website. www.fitbit.com. Accessed: 20.12.2012.
40. Wahoo Fitness website, Wahoo Cycling Speed/Cadence Sensor. www.wahoofitness.com. Accessed: 20.12.2012
41. Skyhook website. www.skyhookwireless.com. Accessed: 20.12.2012
42. V. Manzoni, D. Maniloff, K. Kloeckl, and C. Ratti, "Transportation mode identification and real-time CO2 emission estimation using smartphones," Technical report, Massachusetts Institute of Technology, Cambridge, 2010.
43. Y. Lou, Ch. Zhang, Y. Zheng, X. Xie, W. Wang, and Y. Huang, „Map-matching for low-sampling-rate GPS trajectories," Proc. of 17th ACM SIGSPATIAL Int. Conf. on Advances in Geographic Information Systems (GIS'09), ACM New York, USA, 2009, pp. 352–361.
44. OpenStreetMap website. www.openstreetmap.org. Accessed: 20.12.2012
45. M. Meeuwissen (editor), "Living Lab Plan," EU Project SUNSET, Deliverable D7.1, 2012.
46. K. Lin, A. Kansal, D. Lymberopoulos, and F. Zhao, "Energy-accuracy trade-off for continuous mobile device location," In Proc. of the Int. Conf. on Mobile Systems, Applications and Services (MobiSys'10), ACM, 2010, pp. 285–298.

A Comparison of Automated Keyphrase Extraction Techniques and of Automatic Evaluation vs. Human Evaluation

Richard Hussey, Shirley Williams, Richard Mitchell, Ian Field†
School of Systems Engineering
University of Reading
Reading, United Kingdom
{r.j.hussey, shirley.williams, r.j.mitchell}@reading.ac.uk, IanField90@gmail.com†

Abstract—**Keyphrases are added to documents to help identify the areas of interest they contain. However, in a significant proportion of papers author selected keyphrases are not appropriate for the document they accompany: for instance, they can be classificatory rather than explanatory, or they are not updated when the focus of the paper changes. As such, automated methods for improving the use of keyphrases are needed, and various methods have been published. However, each method was evaluated using a different corpus, typically one relevant to the field of study of the method's authors. This not only makes it difficult to incorporate the useful elements of algorithms in future work, but also makes comparing the results of each method inefficient and ineffective. This paper describes the work undertaken to compare five methods across a common baseline of corpora. The methods chosen were Term Frequency, Inverse Document Frequency, the C-Value, the NC-Value, and a Synonym based approach. These methods were analysed to evaluate performance and quality of results, and to provide a future benchmark. It is shown that Term Frequency and Inverse Document Frequency were the best algorithms, with the Synonym approach following them. Following these findings, a study was undertaken into the value of using human evaluators to judge the outputs. The Synonym method was compared to the original author keyphrases of the Reuters' News Corpus. The findings show that authors of Reuters' news articles provide good keyphrases but that more often than not they do not provide any keyphrases.**

Keywords- Automated Keyphrase Extraction; C-Value; Comparisons; Document Classification; Human Evaluation; Inverse Document Frequency; NC-Value; Reuters News Corpus; Synonyms; Term Frequency

I. INTRODUCTION

The field of natural language processing contains many algorithms devoted to the process of automatic keyphrase extraction (AKE) but the systems lack a common baseline of having been tested on the same corpora.

Previous work by Hussey et al. [1] compared a number of algorithms for AKE, and showed that the best from that set of tests (Term Frequency, Inverse Document Frequency, C-Value, NC-Value, and Synonyms – all explained below) was the statistical method of Term Frequency (listing the terms from the document in order by how often they occurred). However, the same study also laid out areas of further study. This paper sets out to expand on that work, with the expansion of the testing to include the Reuters-21578 corpus and performing a human evaluation of the results.

The original work [1] was based on a study [2] that had shown authors had a tendency to use corpora that were related to or from their own discipline area. For example, those of a medical background used medical corpora (such as the PubMed Central database) while those in literature or linguistics use corpora such as the Journal on Applied Linguistics. This made the task of comparing the effectiveness of one method to another more complex.

Building on the prior work, this study sets out to compare the outputs of all five systems on a set of seven corpora to see if the results of the pilot hold true for a wider range of corpora. The methods chosen are as follows:

- Term Frequency (TF): this ranks words and phrases from the document by how often they occur.
- Term Frequency-Inverse Document Frequency (TD-IDF or Inverse Document Frequency/IDF for short): this also ranks words and phrases from the document by how often they occur, but penalises the rank of any word that also appears frequently in other documents in the same corpus.
- The C-Value [3]: here a series of linguistic filters are used to determine which phrases should be considered, with a ranking metric based on substrings.
- The NC-Value [3]: this follows on from the C-Value, and performs an additional ranking on the outputs of the C-Value – to improve performance.
- The Synonym method [4]: a thesaurus is used to group similar words via their synonyms into keyphrases, which represent common themes of the document.

II. BACKGROUND TO ALGORITHMS

A topic, theme, or subject of a document can be identified by keywords: a collection of words that classify a document. Academic papers make use of them to outline the topics of the paper (such as papers about "metaphor" or "leadership"), books in libraries can be searched by keyword (such as all books on "Stalin" or "romance"), and there are numerous other similar uses. The keywords for a document indicate the major areas of interest within it.

A broader way of capturing a concept is to use a short phrase, typically of one to five words, known as a *keyphrase*. A short phrase of a few linked words can be inferred to

contain more meaning than a single word alone, e.g., the phrase "natural language processing" is more useful than just the word "language".

Sood et al. showed [5] (using the Technorati blog [6] as their source document) that a small number of keywords and keyphrases assigned by humans tend to be used (or reused) frequently. A much larger number of author-supplied keyphrases are idiosyncratic and demonstrate a low frequency as they are too specific to be reused, even by the same author. Examples of reused phrases from Technorati [6] included "politics" and "shopping", while the idiosyncratic phrase examples include "insomnia due to quail wailing". Additionally Sood et al. showed that in half of cases the keyphrases chosen by an author were not suited to the document to which they were attached.

The task faced by AKE is to select the small collection of relevant words that can be used to describe or categorise the document. The process of AKE and its counterpart Automated Keyphrase Assignment (AKA) is discussed by Frank et al. [7]. AKE is characterised by using phrases from the source document (or a reference document) to make the keyphrases. AKA is characterised by using a fixed list of keyphrases and selecting the appropriate ones for the document.

The main aim of this work is to evaluate AKE algorithms for producing keyphrases and to establish a baseline comparison for future studies – as well as to determine which method is best for the corpora used. The secondary aim of this work is to study the usefulness of using human evaluation as opposed to automatic evaluation to determine which is best for ranking algorithms.

The rest of the paper is organised as follows. Section III comprises a review of the algorithms, Section IV results, and Section 0 is a discussion of the outcomes. Section VI then reviews the background of human evaluation, followed by the implementation details in Section VII, and the results are in Sections VIII, IX, X, and XI. Limitations of the study are addressed in Section XII, while Section XIII discusses the results. Section XIVcontains the conclusions of the paper.

III. Review of Algorithms

In this section, relevant methods and the associated results are discussed at a high level. The Term Frequency and Term Frequency-Inverse Document Frequency methods are pure statistical methods, and their generic use is discussed first. Further discussion of the algorithms can be viewed in the original papers [3, 4] as well as in Hussey [8].

While some of the following algorithms are designed with single words in mind, they can be scaled up to include phrases by chunking the text into *n*-grams, as described in Hussey et al. [4, 8].

A. Term Frequency

The "Term Frequency" is simply the number of times a given term (generally a single word) appears in the given document, normalised to prevent bias toward longer documents (longer documents may have higher term counts regardless of importance of the term), as shown in Equation 1. The higher the term frequency, the more likely the term is to be important.

$$tf(t,d) = \frac{f(t)}{n} \quad (1)$$

Where:

- $tf(t,d)$ is the term frequency for term 't' in document 'd'.
- $f(t)$ is the frequency of the occurrence of the term 't' in the corpus.
- n is the number of terms in the document 'd'.

B. Inverse Document Frequency

The "Inverse Document Frequency" is a measure of the importance of the term to the corpus in general terms. This is achieved by dividing the number of documents in the corpus by the number of other documents that contain that term, and then taking the logarithm of the result. This is shown in Equation 2.

$$idf(t) = \log \frac{|D|}{|\{d: t \in d\}|} \quad (2)$$

Where:

- $idf(t)$ is the Inverse Document Frequency for term 't'
- $|D|$ is the total number of documents
- $|\{d: t \in d\}|$ is the number of documents including 't'

Given that if, the term 't' does not occur in the rest of the corpus, the current denominator can lead to a division-by-zero, it is common to alter Equation 2 as shown in Equation 3.

$$idf(t) = \log \frac{|D|}{1+|\{d:t \in d\}|} \quad (3)$$

The IDF is then used as a modifying value upon the term frequency, to reduce the value of those terms that are common across all documents. To achieve this Equation 1 and Equation 3 are combined to form Equation 4.

$$tfidf(t,d) = tf(t,d) \times idf(t) \quad (4)$$

A high weight (indicating importance) is achieved by having a high TF in the given document and a low occurrence in the remaining documents in the corpus – hence filtering out common terms (including stop words such as "the" or "and").

C. C-Value

The C-Value algorithm [3] creates a ranking for potential keyphrases (Frantziy et al. refer to them as "term words") by using the length of the phrase, and the frequency with which it occurs as a sub-string of other phrases.

To start the process, the system tags the corpus with part-of-speech data and extracts strings that pass a linguistic filter (see below) and a frequency threshold. Frantziy et al. used three different linguistic filters (expressed as regular

expressions) in the first stage of the algorithm, and tested the system against each of them. The broader the filter, the more phrases it lets through. Filter 1 is the strictest, whereas Filter 3 is the broadest. The filters were:

1. Noun + Noun
2. (Adj | Noun) + Noun
3. ((Adj | Noun) + | ((Adj | Noun) * (NounPrep)?) (Adj | Noun)*) Noun

Assuming that a phrase a gets through the filter, then its C-Value is calculated as shown in Equation 5. Its value is dependent on whether or not a is a sub-string nested inside another valid phrase.

$$Cvalue(a) = \begin{cases} \log_2|a| \cdot f(a) & a \text{ is not a sub-string} \\ \log_2|a| \cdot \left(f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b) \right) & else \end{cases} \quad (5)$$

Where:

- a is the candidate phrase
- $|a|$ is the length of the phrase a in words
- $f(x)$ is the frequency of the occurrence of 'x'
- T_a is the set of phrases that contain a
- $P(T_a)$ is the number of those phrases

Once the C-Value has been calculated, it is used to rank the phrases and the highest ranked phrases are selected for use as keyphrases.

Frantziy et al. [3] used two metrics to compare the results: Recall and Precision. Recall was the percentage of the keyphrases in the baseline frequency list that were extracted by the C-value algorithm. Precision was the percentage of the keyphrase in the total list that the domain-subject expert agreed with. For Precision, the broader the filter the lower the increase – although all filters showed an improvement of between 1 and 2%. For Recall, the results were broadly similar in tone and dropped the broader the filter from between 2.5% and 2%.

D. NC-Value

The NC-Value [3] extends the C-Value algorithm by using the words adjacent to the keyphrase to add a weighting context to the phrase itself. The weighting is a percentage chance that the word is a context word for a phrase rather than just an adjacent word.

To calculate the NC-Value, the C-Value algorithm is modified by a "*context weighting factor*" which is determined by the nouns, verbs, and adjectives adjoining the keyphrase (these are known as context words). The weight is calculated as shown in Equation 6.

$$weight(w) = \frac{t(w)}{n} \quad (6)$$

Where:

- w is the context word (noun, verb, or adjective)
- $t(w)$ is the number of words 'w' occurs with
- n is the total number of phrases

This is then fed into Equation 7, the NC-Value.

$$NCvalue(a) = 0.8Cvalue(a) + 0.2 \sum_{b \in C_a} f_a(b) weight(b) \quad (7)$$

The values of 0.8 and 0.2 used were arrived at following experimentation by Frantziy et al. [3], and therefore may only be applicable to the medical corpora they used.

Frantziy et al. compared the NC-Value to the C-Value using their previous defined Recall and Precision metrics. The Recall remained the same, as did the average Precision. However the exact Precision varied by section of the output list. The Precision increased in the top section of the list (the top 40 items), and it was reduced in the remainder of the list. This was the expected behaviour, as the aim of the NC-Value was to reorganise the output list to move the better phrases toward the top.

E. Synonyms

The Synonym algorithm [4] takes words from the source document, and groups them together with words that are considered synonyms. It uses a resource document in the form of a thesaurus to aid this. The basic formula for this is shown in Equation 8.

$$KE_N(p_i) = \frac{f(\{w_j : w_j \in S_{p_i}\}) \cdot |w_j|}{|\{S_{p_i}\}|} \quad (8)$$

Where:

- p_i is the candidate phrase
- $f(x)$ is the frequency of occurrence of 'x'
- S_{p_i} is the set of synonyms which p_i belongs to
- $\{w_j : w_j \in S_{p_i}\}$ is all the phrases in set S_{p_i}
- $|w_j|$ is the length of the phrase in words
- $|\{S_{p_i}\}|$ is the number of synonyms in the set

In addition, the unigram list was enhanced by adding the stemmed forms of the unigrams.

However, this method has a tendency to produce a set of keyphrases that are all, almost by definition, synonyms of each other. For example, the words "acquisition" or "taking" can both mean "recovery" [9] and therefore both may have been present as separate keyphrases. To group similar keyphrases into synonym groups, a final step is used. In this step, the algorithm is reapplied to the results of the first application of the algorithm. The aim of this is to prevent a single 'popular' concept from dominating. This involves applying the algorithm again but this time to the generated keyphrases (rather than the document as a whole).

The thesaurus used was Roget's "*Thesaurus of English Words and Phrases*" [9] and the unigrams were stemmed with the Porter Stemming algorithm [10].

F. Baseline

A baseline metric was added to act as a control with which to compare the other algorithms. This algorithm selected words (or phrases) from the source document at random, with a weighting towards shorter phrases. Given the random element it is, therefore, referred to as the Random study.

IV. ALGORITHM RESULTS

This section sets out the results of the five algorithms studied – plus the results of the baseline Random study. The different algorithms were tested against seven corpora. The initial study [1] limited the corpora to those containing academic papers, which for the majority case are submitted with keywords against which the results can be tested.

The initial study had used six corpora: five from the Academics Conferences International (ACI) e-journal [11], and one from the PubMed Central (PMC) database [12]. The ACI papers were on different subject areas:: *Business Research Methods* (EJBRM), *E-Government* (EJEG), *E-Learning* (EJEL), *Information Systems Evaluation* (EJISE), and *Knowledge Management* (EJKM); while the PMC database is an archive of biomedical and life science journal papers.

A seventh, and additional corpus added for this study, was the Reuters-21578 corpus [13] of news articles (which was supplied with keyphrases by the authors of the articles).

The ACI papers [11] were downloaded in August 2009 (when the initial work on this subject was undertaken) and consisted of all the available papers at the time. The selection was not expanded over time, so that later results would remain comparable to earlier results.

The selection of papers from the PubMed Central Corpus [12] were downloaded in their entirety in August 2011. However, as there were 234,496 papers, vastly overshadowing any of the other sources (which tended to average about one hundred papers) a random subset of them was used. To ensure the results retained validity five such samples were taken, and the results averaged over all of them.

For each document analysed in each corpus, the authors had normally supplied an accompanying list of keyphrases to summarise the content. The results of the algorithms were evaluated by comparing them to the author-supplied keyphrases. Where a paper did not have author-supplied keyphrase, it was automatically excluded from the study and the results.

A match was recorded for a paper if at least one of the algorithm keyphrases matched one of the author-supplied keyphrases. The method of comparison was a substring match, which counted two strings as matching if they were equivalent or one of them was a substring of the other. E.g. "know" and "knowledge" would be considered a match. This method was useful for potentially catching instances of keyphrases where the stemmed form or a plural form of the words had been used.

The following tables are all formatted in the same way. They list the 'Corpus' used in the first column and the number of 'Papers' with keyphrases in that corpus. The number 'Matched' is the number of papers that met the above matching criteria as a raw figure and as a percentage (or 'Accuracy' as it is labelled on the tables). The increase (or 'Inc') column, where it occurs, is the numerical value by which the percentage differs from the Random results – i.e. if the match percentage was 1% in the Random study and 10% in the TF study, then that would be an increase of 9.

The results for the C-Value and NC-Value show the range of results over which the three linguistic filters generated outputs – and the Percentages and Increase values are the average for those three results.

All of the results are also summarised below in Figure 1, which can be found after the result tables.

A. Random Study

The Random results showed almost no keyphrases being produced that matched the phrases supplied with the corpora. The results can be seen in Table I.

TABLE I. RANDOM RESULTS

Corpus	Papers	Matched	Accuracy
EJBRM	65	0	0.00%
EJEG	101	2	1.98%
EJEL	111	0	0.00%
EJISE	90	1	1.11%
EJKM	104	5	4.81%
Reuters	21578	3001	13.91%
Reuters-250	216	0	0.00%
Reuters-176	85	0	0.00%
PMC	137	1	0.73%
Average			2.50%

B. Term Frequency

Table II shows the results from the Term Frequency study, and that it performed very well matching on average over 70% of the keyphrases against the authors'.

TABLE II. TF RESULTS

Corpus	Papers	Matched	Accuracy	*Inc*
EJBRM	65	58	89.23%	*89.23*
EJEG	101	93	92.08%	*90.10*
EJEL	111	89	80.18%	*80.18*
EJISE	90	80	88.89%	*87.78*
EJKM	104	101	97.12%	*92.31*
Reuters	21578	3793	17.58%	*3.67*
Reuters-250	216	88	40.74%	*40.74*
Reuters-176	85	66	77.65%	*77.65*
PMC	137	105	76.64%	*75.91*
Average			73.34%	*70.84*

C. Inverse Document Frequency

The Inverse Document algorithm showed a drop in performance compared to the Term Frequency results, as shown in Table III.

TABLE III. TF*IDF RESULTS

Corpus	Papers	Matched	Accuracy	*Inc*
EJBRM	65	43	66.15%	*66.15*
EJEG	101	66	65.35%	*63.37*
EJEL	111	69	62.16%	*62.16*
EJISE	90	69	76.67%	*75.56*
EJKM	104	71	68.27%	*63.46*
Reuters	21578	1748	8.10%	*-5.81*
Reuters-250	216	13	6.02%	*6.02*
Reuters-176	85	35	41.18%	*41.18*
PMC	137	107	78.10%	*77.37*
Average			52.44%	*49.94*

D. The C-Value

As there were three linguistic filters for the C-Value, the results in Table IV show the range of the matched values and then an averaged percentage.

TABLE IV. C-VALUE RESULTS

Corpus	Papers	Matched	Accuracy	*Inc*
EJBRM	65	10-19	~23.08%	*~23.08*
EJEG	101	16-30	~23.76%	*~21.78*
EJEL	111	1-5	~1.80%	*~1.80*
EJISE	90	11-12	~12.22%	*~11.11*
EJKM	104	3-7	~4.81%	*~0.00*
Reuters	21578	87-145	~0.54%	*~-13.37*
Reuters-250	216	0-1	~0.46%	*~0.46*
Reuters-176	85	2-4	~3.53%	*~3.53*
PMC	137	25-31	~21.17%	*~20.44*
Average			~10.15%	*~13.03*

E. The NC-Value

Similar to the C-Value, the results for the NC-Value are displayed as ranges for the matches and as an average percentage.

TABLE V. NC-VALUE RESULTS

Corpus	Papers	Matched	Accuracy	*Inc*
EJBRM	65	1-4	~3.08%	*~3.08*
EJEG	101	0	0.00%	*-1.98*
EJEL	111	0	0.00%	*0.00*
EJISE	90	0	0.00%	*-1.11*
EJKM	104	0	0.00%	*-4.81*
Reuters	21578	0-1	~0.00%	*~-13.91*
Reuters-250	216	0	0.00%	*0.00*
Reuters-176	85	0	0.00%	*0.00*
PMC	137	0	0.00%	*-0.73*
Average			~0.34%	*~-2.16*

F. Synonym Study

The synonym results show a good improvement over the baseline results (nearly 50% on average), although particular corpora fared poorly (the medical corpus PMC for example, compared to the Knowledge Management corpus). The results are shown in Table VI.

TABLE VI. SYNONYM RESULTS

Corpus	Papers	Matched	Accuracy	*Inc*
EJBRM	65	31	47.69%	*47.69*
EJEG	101	73	72.28%	*70.30*
EJEL	111	77	69.37%	*69.37*
EJISE	90	46	51.11%	*50.00*
EJKM	104	94	90.38%	*85.57*
Reuters	21578	1040	4.82%	*-9.09*
Reuters-250	216	26	12.04%	*12.04*
Reuters-176	85	43	50.59%	*50.59*
PMC	137	47	34.31%	*33.58*
Average			48.07%	*45.56*

V. ALGORITHM DISCUSSION ON RESULTS

The results outlined in Section IV above show that the Term Frequency algorithm had the highest percentage matches of any of the algorithms. Figure 1, below, groups and summarised these results.

The keyphrases supplied by authors are always likely to contain at least one "common" word that would show up in a frequency count. This would also explain the poor results produced by Inverse Document Frequency algorithm, as common words in the corpus are likewise likely to be keyphrases supplied by the author. For example, the papers in EJKM use on average the phrase "knowledge" 102 times per paper (11,675 times over 114 papers) and only 15 papers do not include it as an author supplied keyword. Therefore, there is a high likelihood that a count of word frequencies will select this as one of the five keyphrases from the TF algorithm. Due to this proliferation across the corpus, this would also explain its absence from the TF*IDF results (as TF*IDF ranks words which are common between documents as less important) and, therefore, the lower number of papers where a match was recorded.

The C-Value [3] was not predicted to perform as poorly as it did, given that the paper the algorithm was taken from reported Precision of approximately 30% (across all three filters) while Recall for all three systems was at nearly 100%.: these values were the same for the NC-Value as well. Furthermore, the SNC-Value [14] successfully built on the results of the NC-Value.

It is clear from the above results that in all likelihood an error occurred in the implementation of the algorithm, but despite multiple attempts to locate a difference between the published algorithm and the implemented code, none could be found at the time of writing.

Sood et al. [5] showed that keyphrases chosen by the authors of documents are chosen inappropriately 51.15% of the time. These factors combined suggest that the matching criteria should be changed for future work and a

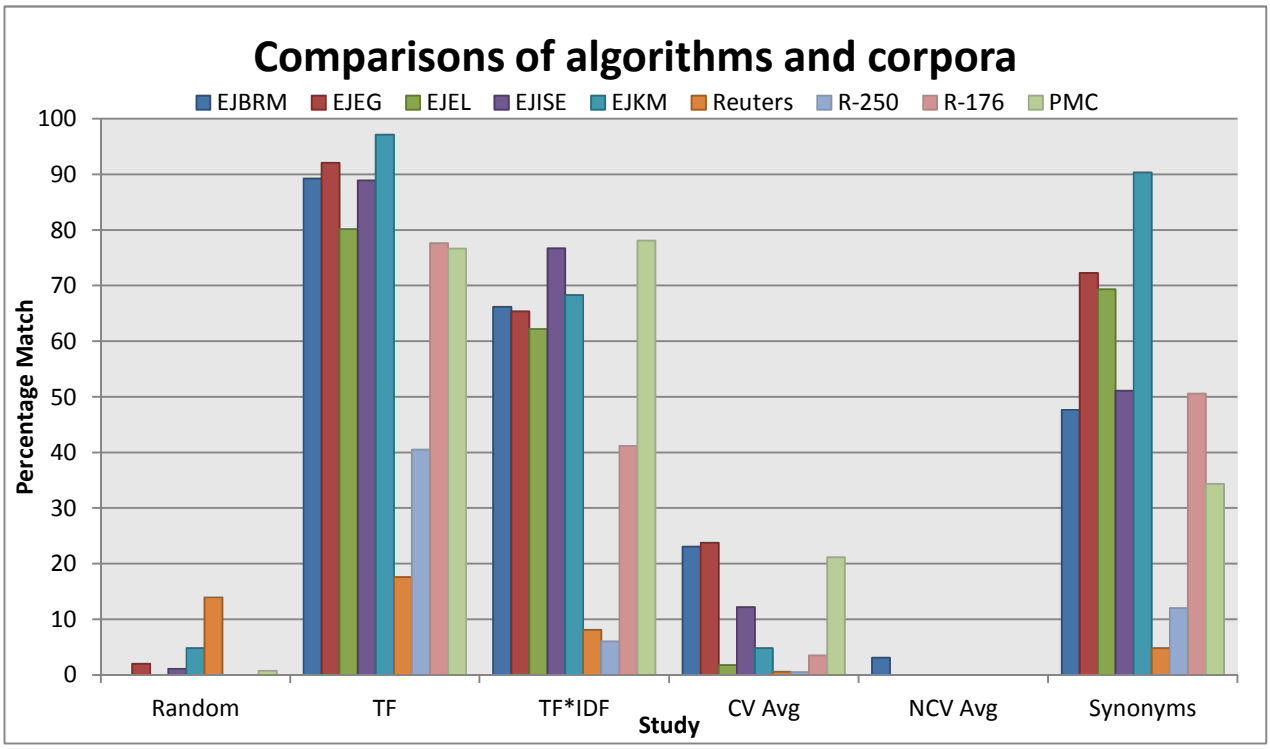


Figure 1. Algorithms Matches by Corpus

Recall/Precision model, as used by Frantzi et al., would seem appropriate.

In conclusion, it can be stated that the results of the study show that when using the naïve comparison method, the results are biased towards phrases that occur most often in the document. Later sections will look at the results of implementing Recall/Precision and testing other corpora and evaluation metrics (such as using human judges to compare the Synonym method [4] to the Reuters' News corpus [13]).

VI. Background to Human Evaluation

In the literature, various other authors also used human judges to evaluate the results of AKE techniques. For example, the keyword extraction algorithm developed by Matsuo & Ishizuka [15] was tested by 20 human judges who were presented with their own published papers, and asked to pick any number of appropriate keywords from a list. The list was populated with the top keywords from Matsuo & Isnizuka's algorithm, from KeyGraph [16], and the statistical measure TF and TF-IDF. The top 15 keyphrases from each of the systems were shuffled into a list, which was presented to the judges. The judges were then asked to pick whichever keyphrases they felt were appropriate, and in addition to select five that they felt were "indispensible" for describing the paper. These indispensible terms could be supplemented by terms from the paper, if the judges felt that systems had not generated appropriate candidates. The results were assessed by Coverage (ratio of indispensible terms in the generated keywords to the number of indispensible terms), Frequency Index (average frequency of the terms in the list from each system), and Precision (ratio of terms chosen by judges to the number of terms generated by that system). Matsuo & Isnizuka's system outperformed the alternatives on Coverage and Frequency Index, but TF and IDF performed better at Precision.

Sood et al [5] used a panel of ten judges to evaluate their system for tagging blog posts, TagAssist. Blog post data from Technorati [6] was analysed by TagAssist to produce keyphrases. The judging panel were presented with author-chosen keyphrases, the TagAssist keyphrases, and the keyphrases generated by a baseline comparison system. Without being aware of the source algorithm of each of the keyphrases, the judges were asked to pick those that they felt thought suitable for describing the associated blog post. The results of the human judges found that, while the original author keywords were the best, in over half the cases (51.15%) they were not appropriate. After the original author tags (48.85%), those produced by TagAssist were ranked second (42.10%), followed by the baseline (30.05%).

Barker and Cornacchia presented the union of the sets of keyphrases from Extractor [17] and their own system [18] to twelve human judges to place in three categories: "Good", "So-so", and "Bad" which were then converted into a 'points' value of 2, 1, and 0 respectively. The B&C set of keyphrases averaged a lower score (0.47) than Extractor (0.56), but the difference was not statistically significant. The two sets of keyphrases were also presented to the judges unmerged, and they were asked to pick which set better represented the document, or if neither represented it particularly well. The results of this part of the testing

showed that the judges thought the B&C keyphrases were better 47% of the time, and the Extractor set 39% of the time. However, when the significance of the judge's decisions was analysed (through use of the Kappa Statistic [19]) it was no greater than if they had selected response at random – and therefore the results of this study are not particularly fruitful, with statistically insignificant differences and only chance variations of results.

Frantziy et al. tested the C-Value [3] with a domain expert (in the area of eye pathology medical records, which was their test corpus) who was presented with the terms extracted by the system and asked to indicate which they agreed with. The number in agreement as a fraction of the whole was classed as the Precision of the system. The authors also calculated Recall as the fraction of the baseline list of terms that the C-Value also selected. The extension to the system, NC-Value, was tested in the same fashion and showed no change in the Recall but had a different Precision depending on which section of the output list was compared – this was the expected behaviour as the NC-Value did not create any additional terms but simply attempted to reorder them to get a better fit. In both systems, Recall was improved by using stricter and narrower linguistic filters. The SNC-Value (also called 'TRUCKS') is a further extension of the NC-Value which was also tested by domain experts, but this time [14] used two such experts rather than one although their involvement in the evaluation process was kept the same.

A. *Precision, Recall, and Harmonic Mean*

As discussed above, Precision and Recall are measures often used in Information Retrieval and AKE for comparing the outputs of different systems – and determining how close to a 'perfect' or gold-standard system they have become. Several of the papers referenced here have made use of these measures, these are summarised below:

- Precision [3] – the number of terms in agreement with the whole
- Precision [15] – the ratio of terms chosen by judges to the number of terms generated by that system
- Precision [20] – the ratio of relevant sentences in the summary to the number of sentences in the summary
- Precision [21] – the ratio of relevant sentences in the summary to the number of sentences in the summary
- Precision [22] – the fraction of relevant keywords compared to the whole
- Precision [23] – the fraction of extracted sentences also in the model summary
- Precision [24] – Number of correctly predicted keyphrases divided by the total number of predictions.
- Recall [3] – the fraction of the baseline also selected by C-Value
- Recall [20] – the fraction of the relevant sentences in the document that were also in the summary
- Recall [21] – the fraction of the relevant sentences in the document that were also in the summary
- Recall [22] – the fraction of relevant keywords compared to the total relevant
- Recall [23] – the ratio of extracted sentences in the model summary to the number of sentences in the model summary
- Recall [24] – total number of correctly predicted keyphrases divided by the number of 'gold standard keyphrases' (the keyphrases supplied by the authors of the considered papers).

Based on the above, generic definitions of Precision and Recall can be deduced as follows:

- Precision is most often expressed as the number of useful phrases generated by the system divided by the total number of phrases generated. In this paper, Precision is defined as the number of phrases chosen by the judges that were also phrases supplied by Reuters, divided by total number of phrases generated for that system. See also Equation 9.
- Recall is most often expressed as the number of useful keyphrases generated by the system divided by the number of keyphrases in the baseline, or 'gold-standard', system. In this paper, Recall is defined as the number of phrases chosen by the judges that were also phrases supplied by Reuters, divided by the number of phrases supplied by Reuters. See also Equation 10.

$$Precision(System) = \frac{Cooccurrence(System\ with\ Reuters)}{Possible\ Selections\ (System)} \quad (9)$$

$$Recall(System) = \frac{Cooccurrence(System\ with\ Reuters)}{Possible\ Selections\ (Reuters)} \quad (10)$$

The Harmonic Mean of Precision and Recall (also known as the F-Measure or F-score) is also used by Joshi & Matwani [22], Jones et al [21]., Lin & Hovy [23], and Goldstein et al [20]. The harmonic mean is a variant mean, which is different to the 'normal' (arithmetic) mean, and its generic equation for variables $x_1 \cdots x_n$ is shown in Equation 11. In addition, the special case of the Harmonic Mean for x_1 and x_2 is shown in Equation 12.

$$H = \frac{n}{\sum_{i=1}^{1n} \frac{1}{x_i}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} \quad (11)$$

$$H = \frac{2 \times x_1 \times x_2}{x_1 + x_2} \quad (12)$$

As the Harmonic Mean calculated in this paper will be for the values of Precision and Recall, the special case shown in Equation 12 will be used.

VII. HUMAN EVALUATION

A system was created to test the outputs of the Synonym algorithm on the Reuters' News Corpus [13], a commonly used corpus in the natural language processing community, consisting of 21,578 short news articles from the Reuters news network. An example is given in Figure 2; however, as shown in Figure 3, not all of the entries in the corpus make immediate sense as they often written in a form of shorthand.

The Reuters' corpus also contains keyphrases for the articles, as supplied by the authors of the articles. Therefore, the chosen method of assessing the results of the system was to present the judges with a list comprising of five keyphrases from the Reuters' articles, five keyphrases from the author's algorithm, and five keyphrases chosen at random from the news article. The keyphrases were then sorted alphabetically, and shown to the judge. The judges had no specific domain knowledge relevant to the task. They were recruited by accessing the link to the website, which was e-mailed to the authors' institution as well as promoted on popular social networking sites.

The judge was asked to select any appropriate keyphrases or indicate that none of them was suitable. The articles shown to the judge were selected at random from the whole set of viable articles, and the judges were asked to evaluate as many articles as they desired. For the first experiment, the viable articles were all 21,578 articles, and for the second the viable articles were the 125 articles that had five or more author-keyphrases.

For example for article 69 the website would show the article text (shown in Figure 2), followed by the alphabetical listing of the associated keyphrases (shown in Figure 4). Fifteen keyphrases are shown: five from the original Reuters article, five from the synonym analysis, and five chosen from the text by chance. The same keyphrases are shown again in Figure 5, but separated out into their original groups. Figure 6 shows the finished layout of this information on the website used to capture the data.

To ensure the widest distribution of the test, the testing was done via a simple website which displayed the article number, the article text, the possible keyphrases, and a submit button. When the web page was loaded, an article was selected at random. The user clicked the submit button to insert the data into the database, and then the page was refreshed automatically – presenting the user with a new article.

VIII. RESULTS FOR ALL ARTICLES

The first study undertaken was with all 21,578 articles of the Reuters' corpus.

A. Evaluation and Results

The first iteration of the testing tool displayed one of the 21,578 articles to the user, and generated the results shown in Table VII. There were 250 submissions for this test.

Reporting members of the National Soybean Processors Association (NSPA) crushed 21,782,929 bushels of soybeans in the week ended Feb 25 compared with 22,345,718 bushels in the previous week and 16,568,000 in the year-ago week, the association said. It said total crushing capacity for members was 25,873,904 bushels vs. 25,873,904 last week and 25,459,238 bushels last year. NSPA also said U.S. soybean meal exports in the week were 117,866 tonnes vs. 121,168 tonnes a week ago and compared with 84,250 tonnes in the year-ago week. NSPA said the figures include only NSPA member firms. Reuter

Figure 2. Reuters 21578 corpus #69

Six months to December 31 shr 8.8 cts vs. 0.5 ct interim dividend 12.5 cts vs. nil group net 9.5 mln ringgit vs. 0.6 mln pre-tax 11 mln vs. 1.1 mln turnover 88.9 mln vs. 70.8 mln note - dividend pay may 15, register April 17. Reuter

Figure 3. Reuters 21578 corpus #2962

"345 718 bushels", "568 000", "calefaction", "crushed 21 782", "for members was 25", "fuel", "lubrication", "meal-feed", "oil", "oilseed", "remedy", "soy-meal", "soybean", "total crushing capacity for", "veg-oil"

Figure 4. Alphabetical keyphrases for article #69

Reuters - "veg-oil", "soybean", "oilseed", "meal-feed", "soy-meal"

Synonym - "oil", "fuel", "lubrication", "calefaction", "remedy"

Chance - "total crushing capacity for", "for members was 25", "568 000", "crushed 21 782", "345 718 bushels"

Figure 5. Seperated keyphrases for article #69

TABLE VII. RESULTS OF FIRST TEST

Method	Total Keyphrases	Selected Keyphrases	Mean Selected
Original Reuters	14,058	42	0.16800
Synonyms	107,225	146	0.58400
Random	107,885	178	**0.71200**
Total	229,838	366	0.48800

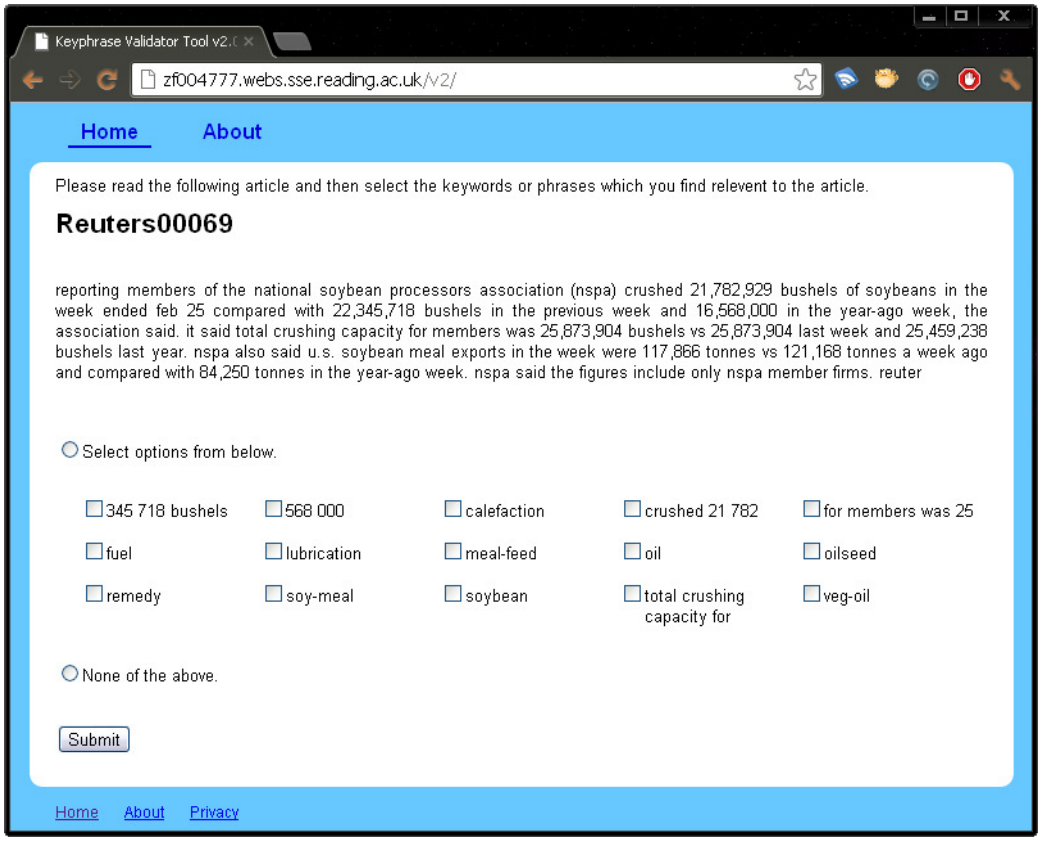


Figure 6. Website layout for article #69

For the column "Total Keyphrases", the number displayed is the number of keyphrases for each article (21,578 articles) multiplied by the number of articles with that number of keyphrases. Therefore, for the original Reuters' keyphrases there were 125 articles with five or more keyphrases, giving 625 keyphrases for that subsection. This was calculation repeated for the number of keyphrases between 0 and 5 (as only five keyphrases were stored per article all articles with more than five keyphrases were treated as having five) – the values for each of these are shown in Table VIII.

TABLE VIII. NUMBER OF ARTICLES WITH X NUMBER OF KEYPHRASES

Number of Keyphrases	Number of Articles	Percentage of Corpus	Total Keyphrases
0	10,273	47.60%	0
1	9,443	43.76%	9,443
2	1,324	6.14%	2,648
3	301	1.39%	930
4	103	0.48%	415
5+	125	0.58%	625
Total	21,578		14,058

For the Synonyms and Random, there were always five keyphrases so the value in the "Total Keyphrases" column is simply 21,578 times five (107,890).

The "Selected Keyphrases" column shows the number of those keyphrases which were picked in total over all of the 250 user submissions. The "Mean Selected" column shows the average number of keyphrases chosen per submission for that category of keyphrase.

As can be seen from the Mean values, in the average case less than one keyphrase was selected by the users and the keyphrases chosen at random were selected more often than the Synonyms or the original Reuters' keyphrases.

Table IX shows the data from Table VII combined with the number of selections made as well as the total possible selections – which was calculated from the number of keyphrases available in each of the 250 responses. Therefore, for Synonyms and Random, this was 250 times five, whereas for the Reuters there were only 164 across the entries.

TABLE IX. MATCHES AND SELECTION FOR FIRST TEST

	Reuters	Synonyms	Random
Selected	42	146	178
Submissions	36	106	109
Possible Selections	164	1250	1250
Mean Matches	0.16800	0.58400	0.71200

In order to calculate the Recall, the keyphrase co-occurrence needed to be calculated. This is the number of keyphrases in one system that also appeared in each of the

other systems. This was done by a strict matching policy that only recorded a match if both strings were equal. The results of this are shown in Table X.

TABLE X. KEYPHRASE CO-OCCURRENCE MATRIX FOR FIRST TEST

	Reuters	Synonyms	Random
Co-occurrence with Reuters	42	5	0
Co- occurrence with Synonyms	3	146	0
Co- occurrence with Random	12	36	178
Total co-occurrences	57	187	178

From Table IX and X the Precision and Recall of the two systems and the original keyphrases can be calculated. This is set out, below in Equations 13 to 18.

$$P(Reuters) = \frac{42}{164} = 0.2561 \quad (13)$$

$$P(Synonyms) = \frac{5}{1250} = 0.0040 \quad (14)$$

$$P(Random) = \frac{0}{1250} = 0 \quad (15)$$

$$R(Reuters) = \frac{42}{164} = 0.2561 \quad (16)$$

$$R(Synonyms) = \frac{5}{164} = 0.0305 \quad (17)$$

$$R(Random) = \frac{0}{164} = 0 \quad (18)$$

The Harmonic Mean of the three systems is shown in Equations 19 to 21 (using the special case of the Harmonic Mean formula for two values, as shown in Equation 12).

$$H(Reuters) = \frac{2 \times 0.2561 \times 0.2561}{0.2561 + 0.2561} = 0.2561 \quad (19)$$

$$H(Synonyms) = \frac{2 \times 0.0040 \times 0.0305}{0.0040 + 0.0305} = 0.0071 \quad (20)$$

$$H(Chance) = \frac{2 \times 0 \times 0}{0 + 0} = 0 \quad (21)$$

Collating all these results, and including the Harmonic Mean, gives the results shown in Table XI below. As can be seen from the table, the Reuters' keyphrases came out top on all three measures – despite the fact that the Random keyphrases had a better selection mean (see Table VII). This is because the Reuters' keyphrases were deemed (by the Precision and Recall) to be a better fit to the baseline, rather than simply selected more often. Section X discusses in more detail the suitability of the selected keyphrases from all three sets.

TABLE XI. PRECISION, RECALL, AND HARMONIC MEAN VALUES FOR FIRST TEST

	Reuters	Synonyms	Random
Precision	**0.2561**	0.0040	0
Recall	**0.2561**	0.0305	0
Harmonic Mean	**0.2561**	0.0071	0

B. Statistical Significance

To ensure that there was a statistically significant difference between the results of the different methods, the one-way ANOVA process was run on the submissions from the website (the 250 submissions from all possible articles). ANOVA stands for **An**alysis **o**f **Va**riance, and uses a probability distribution (F-distribution) with information about the variance of the populations ('within' samples) and the grouping of the populations ('between' samples) to determine if the difference between and within the populations are actually different or could have arisen from chance [25]. Expressed another way; the ANOVA calculation tests the hypotheses shown in Equation 22 to see if the null hypothesis (H_0) can be rejected in favour of the alternative (H_a). The null hypothesis is that the means of the results are the same, and that any variance in the results is due to the perturbations of chance.

$$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k \; or \qquad H_a: \exists \mu_m, \mu_n (\mu_m \neq \mu_n) \quad (22)$$

The results of the ANOVA table are shown in Table XII.

TABLE XII. ANOVA TABLE FOR FIRST TEST

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-Statistic	P-Value
Between Samples	40.4480	2	20.2240	33.0617	1.75×10^{-14}
Within Samples	456.9440	747	0.6117		
Totals	497.3920	749			

The P-Value is the statistical likelihood that the results gained were found by chance. Normally, such statistics are evaluated at the 95% confidence level (P-Value of 0.05) or the 99% confidence level (P-Value of 0.01) – which means that there is a 5% (or 1%) probability that the results arose because of chance, and therefore a 95% (or 99%) probability the results are statistically valid.

As can be seen from Table XII, the calculated P-Value for this data is:

$$\text{P-Value} = 1.75 \times 10^{-14}$$

This means that the null hypothesis can be rejected with over a 99% confidence level.

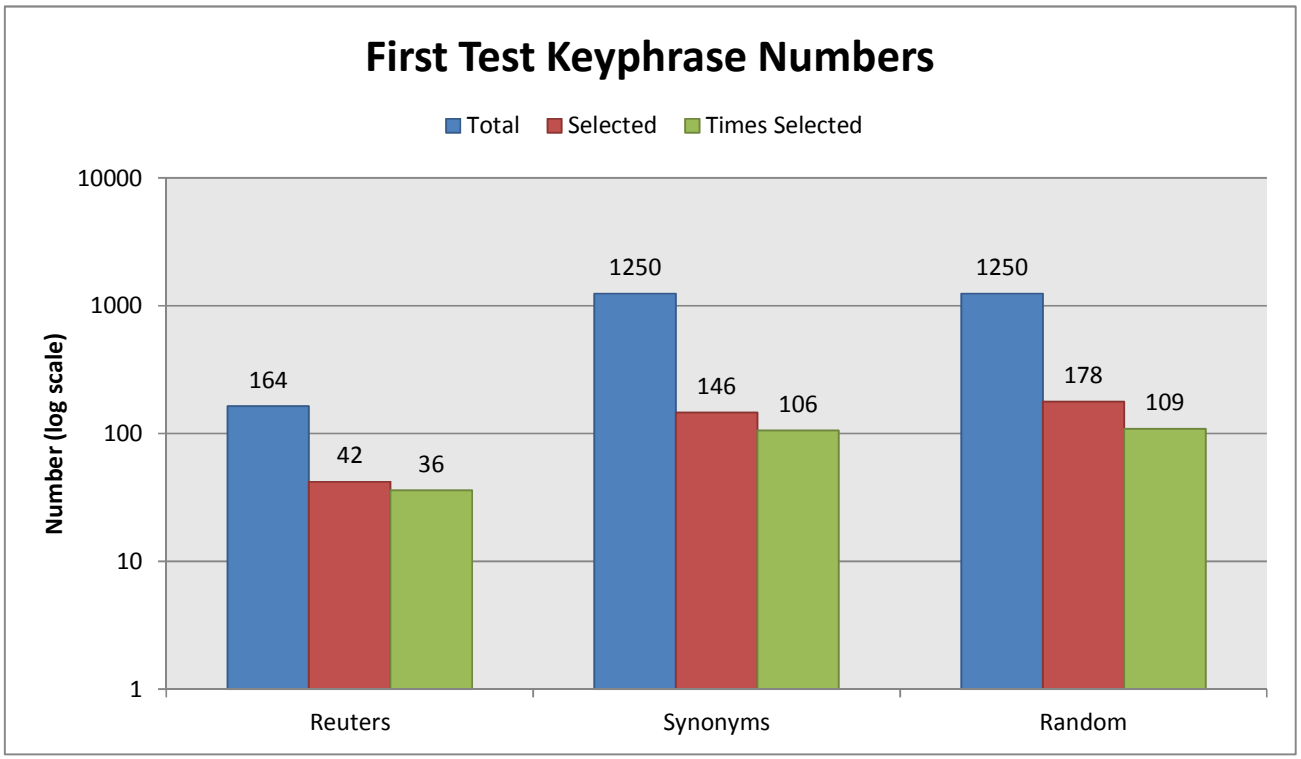


Figure 7. Graph of First Test Keyphrases (Log scale)

C. Discussion

The first test showed that the best system was the Random selection; 178 of the keyphrases picked (48.6%) were supplied by it. Following this was the keyphrases from the synonym system (146 out of 366, or 39.9%), and then the original keyphrases supplied by Reuters (11.5%).

Figure 7 shows these values plotted side-by-side, and because the total number of keyphrases was sufficiently large compared to the number actually selected over the test the graph is plotted on a $\log_{10}$-scale so that the values can be seen together.

However, the number of keyphrases for the Synonyms and Random options were over seven times larger than the Reuters' keyphrases (which, in turn, only made up 6% of the total 229,838 keyphrases in the system). Therefore, there was a clear chance that for any given article there would be more keyphrases chosen by the algorithms than came with the article. Indeed, 48% of the articles (10,273 out of 21,578) had no keyphrases supplied by Reuters.

Therefore, it was decided to run the test again, but only displaying the articles that had at least five keyphrases from Reuters.

IX. Results for 5+ Keyphrases

Following analysis of the results from Section VIII, a second version of the site were created which only displayed those articles which had at least five keyphrases associated with them.

A. Evaluation and Results

Displaying only these 125 articles to the user generated the results shown in Table XIII. There were 176 submissions for this test.

TABLE XIII. Results of the Second Test

Method	Total Keyphrases	Selected Keyphrases	Mean Selected
Original Reuters	625	324	**1.84091**
Synonyms	625	81	0.46023
Random	625	95	0.53977
Total	1,875	500	0.94677

This time, the column "Total Keyphrases" has the same number of keyphrases for each method as the Synonym and Random algorithms always output five keyphrases, and as such only ever five keyphrases from the Reuters' list were ever shown. The "Selected Keyphrases" column, again, shows the number of those keyphrases which were picked in total over all of the 176 user submissions. The "Mean Selected" column shows the average number of keyphrases chosen per submission for that category of keyphrase.

As can be seen from the Mean values in Table XIII, the average case was better than in Table VII– nearly two times the value. While the average did not quite reach one keyphrase per selection, the Reuters' results nearly averaged two keyphrases per selection and the Random keyphrases outperformed the Synonym results.

From Table XIV and Table XV (Keyphrase Co-occurrence) the Precision, Recall, and Harmonic Mean of the three systems can be calculated. This is set out, below in Equations 22 to 30.

TABLE XIV. MATCHES AND SELECTIONS FOR SECOND TEST

	Reuters	**Synonyms**	**Random**
Matches	324	81	95
Entries	625	625	625
Selections	136	64	60
Possible Selections	880	880	880
Mean Matches	1.84091	0.46023	0.53977

TABLE XV. KEYPHRASE CO-OCCURRENCE MATRIX FOR SECOND TEST

	Reuters	**Synonyms**	**Random**
Co-occurrence with Reuters	324	8	0
Co- occurrence with Synonyms	3	81	0
Co- occurrence with Random	76	22	95
Total co-occurrences	403	111	95

$$P(Reuters) = \frac{324}{880} = 0.3682 \quad (22)$$

$$P(Synonyms) = \frac{8}{880} = 0.0091 \quad (23)$$

$$P(Random) = \frac{0}{625} = 0 \quad (24)$$

$$R(Reuters) = \frac{324}{880} = 0.3682 \quad (25)$$

$$R(Synonyms) = \frac{8}{880} = 0.0091 \quad (26)$$

$$R(Random) = \frac{0}{880} = 0 \quad (27)$$

$$H(Reuters) = \frac{2 \times 0.3682 \times 0.3682}{0.3682 + 0.3682} = 0.3682 \quad (28)$$

$$H(Synonyms) = \frac{2 \times 0.0091 \times 0.0091}{0.0091 + 0.0091} = 0.0091 \quad (29)$$

$$H(Chamce) = \frac{2 \times 0 \times 0}{0 + 0} = 0 \quad (30)$$

Collating all these results, and including the Harmonic Mean, gives the results shown in Table XVI below. As can be seen from the table, again the Reuters' keyphrases came out top on all three measures – which in this test coincided with the Reuters' keyphrases having the better selection mean (see Table XIII).

TABLE XVI. PRECISION, RECALL, AND HARMONIC MEAN VALUES FOR SECOND TEST

	Reuters	**Synonyms**	**Random**
Precision	**0.3682**	0.0091	0
Recall	**0.3682**	0.0091	0
Harmonic Mean	**0.3682**	0.0091	0

B. Statistical Significance

To ensure that the results were statistically significant the one-way ANOVA process was run on this set of submissions (the 176 submissions for the articles with 5+ keyphrase) as well, and the results are shown in Table XVII.

TABLE XVII. ANOVA TABLE FOR SECOND TEST

Source	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	**F-Statistic**	**P-Value**
Between Samples	211.5265	2	105.7633	86.3557	3.78×10^{-33}
Within Samples	642.9886	525	1.2247		
Totals	854.5152	527			

Once again, the P-value for the likelihood of the results arising due to chance was much smaller than required to prove the results statistically significant, even smaller than the results from the first test. As the P-Value shows, the results are statistically significant.

C. Discussion

The second test showed a reversal of the results from the first test. The 'best' system was where the keyphrases were supplied by Reuters, as 324 of the keyphrases picked (64.8%) were from this source. Following this was the random keyphrases selection (95 out of 500, or 19.0%), and then the synonym keyphrases (16.2%). Figure 8 shows these values plotted side-by-side, and to remain consistent with Figure 7 the test the graph is plotted on a $\log_{10}$-scale so that the values can be seen together.

Due to the changes made for this test, the number of keyphrases was consistent across each source (625 per source). However, it was noted that while more keyphrases were selected in total from the Random source, they were selected fewer times. The 95 selected Random keyphrases were picked over only 60 entries of the 176-recorded entries (34.1%) while the 81 Synonym keyphrases were picked in 64 of the entries (36.4%). This is at odds with the first test, where the number of selections from each source was proportional to the number of keyphrases selected overall.

Again, the Precision and Recall measures, and the Harmonic Mean, showed that Reuters' keyphrases were best at representing the gold standard.

X. INDIVIDUAL ENTRIES

To examine further this outcome, some of the individual results are discussed below. The results chosen were selected because they all shared the same base article being judged – thus allowing a comparison to be drawn.

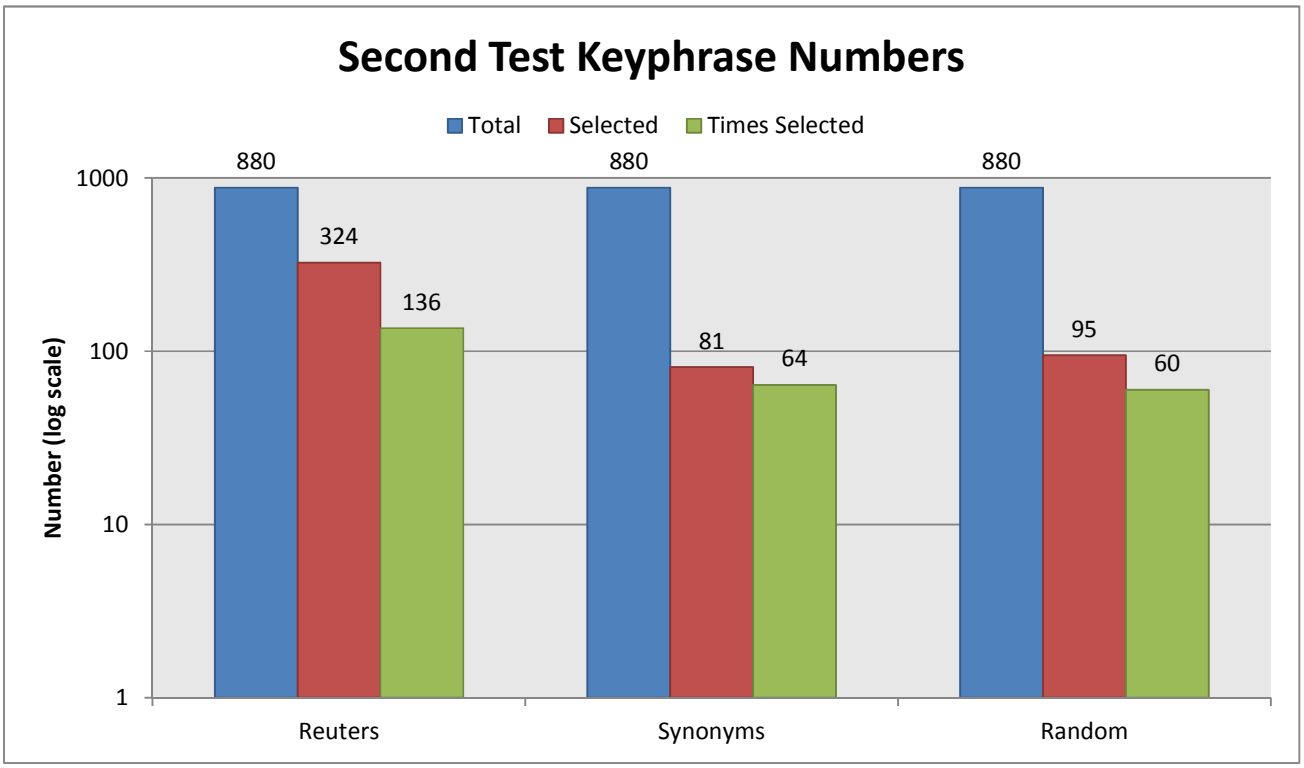


Figure 8. Graph of Second Test Keyphrase Numbers (Log scale)

Table XVIII shows the entries submitted for article number 69 (the text of which was shown above in Figure 2). The results for these entries show the same distribution of keyphrases on average as the overall results.

TABLE XVIII. RESULTS OF REUTERS ARTICLE #69

System	Keyphrase	#1	#2	#3	Sum	Avg
Reuters'	Veg-Oil		1			
	Soybean	1	1	1		
	Oilseed					
	Meal-Feed					
	Soy-Meal			1	**5**	**1.667**
Synonyms	Oil					
	Fuel					
	Lubrication					
	Calefaction					
	Remedy				0	0.000
Random	Total Crushing Capacity		1			
	For					
	For Members was 25					
	586 000					
	Crushed 21 782		1			
	345 718 Bushels				2	0.667
Total		1	4	2	7	2.333

The results show that certain keyphrases from the original set were never picked – as they do not seem to be relevant (which fits with the expectations gained from the literature [5], that only a certain percentage of user-supplied keywords are appropriate). However, the selected Random keyphrases seemed, while related to the article, unsuitable as keywords or phrases. See, for example, "Total Crushing Capacity For".

Table XIX shows the collected entries for article number 10,172 (the text of which is in Figure 9). The first seven entries were taken from the second test, but the eighth was the submission for the same article in the first test.

TABLE XIX. RESULTS OF REUTERS ARTICLE #10,172

System	Keyphrase	1	2	3	4	5	6	7	*8*	Sum	Avg
Reuters'	Grain	1	1	1			1		*1*		
	Corn	1				1	1	1	*1*		
	Wheat	1					1	1	*1*		
	Oilseed										
	Soybean	1	1	1	1	1	1	1	*1*	**22**	**2.750**
Synonyms	Agreement	1			1		1				
	Concord										
	Assent										
	Uniformity										
	Expedience									3	0.375
Random	Third Year										
	of the U S										
	The Fourth										
	Year										
	The Soviet		1	1							
	Which					1					
	Ended										
	To the U S									3	0.375
Total		5	3	3	2	3	5	3	*4*	28	3.500

Again, the distribution of keyphrases per source was in line with the overall averages – with Reuters having the bulk of the selections (78.6%). The selected Random keyphrases this time were more, subjectively, appropriate for the article and better fits for keyphrases.

There were no shipments of U.S. grain or soybeans to the Soviet Union in the week ended March 19, according to the U.S. agriculture department's latest export sales report. The USSR has purchased 2.40 mln tonnes of U.S. corn for delivery in the fourth year of the U.S.-USSR grain agreement. Total shipments in the third year of the U.S.-USSR grains agreement, which ended September 30, amounted to 152,600 tonnes of wheat, 6,808,100 tonnes of corn and 1,518,700 tonnes of soybeans. Reuter

Figure 9. Reuters-21578 corpus #10,172

XI. DATA CLEANSING

Following the insights gained from studying the individual results of testing it was concluded that the data acquired in the second test (Section IX) required sanitising/cleansing to remove entries that clearly had no use as keyphrases. Some of the keyphrases selected by the judges were not useful keyphrases, so they were removed from consideration. Examples included "Which Ended", "Total Crushing Capacity For", or even "940 < title>blah blah".

Therefore, the data from the articles with five+ keyphrases (detailed in Section IX above) was cleansed to remove such non-useful keyphrases, and the same calculations were run to determine what changes this produced.

Random keyphrases were marked as not selected if the phrase did not constitute a valid phrase linguistically speaking – which is to say a single unit in the syntax of a sentence. The articles were not read for this process, so any selection that was a valid phrase was retained regardless of whether it was appropriate. This decision was taken as it was the purpose of the website to capture which valid phrases were appropriate for the article and to take that step in the data cleansing would have invalidated all of the judges' work.

A. Evaluation and Results

The results for the cleansed data are shown in Table XX. There remained 176 submissions over 125 articles.

TABLE XX. RESULTS OF DATA CLEANSING

Method	Total Keyphrases	Selected Keyphrases	Mean Selected
Original Reuters	625	324	**1.84091**
Synonyms	625	81	0.46023
Random	625	23	0.13068
Total	1,875	428	0.81061

As can be seen from the Mean values in Table XX, the average case has dropped from Table XIII – as was expected as the Random mean has dropped substantially and is now below the value for the Synonym keyphrases.

Once more, from Table XXI and XXII the Precision and Recall of the three systems can be calculated. However, as only the Random data has been altered, the values for Reuters and Synonyms stay the same and the following equations list the new values for Random –this is shown in Equation 31 and Equation 32. Again, Table XXI shows the Matches, Entries, and the Selections numbers – calculated as before.

TABLE XXI. MATCHES AND SELECTION FOR DATA CLEANSING

	Reuters	Synonyms	Random
Selected	324	81	95
Submissions	136	64	60
Possible Selections	880	880	880
Mean Matches	1.84091	0.46023	0.13068

TABLE XXII. KEYPHRASE CO-OCCURRENCE MATRIX FOR DATA CLEANSING

	Reuters	Synonyms	Random
Co-occurrence with Reuters	324	8	0
Co- occurrence with Synonyms	3	81	0
Co- occurrence with Random	76	22	23
Total co-occurrences	403	111	23

$$P(Chance) = \frac{0}{880} = 0 \quad (31)$$

$$R(Chance) = \frac{0}{880} = 0 \quad (32)$$

The Harmonic Mean is then calculated in Equation 33.

$$H(Chance) = \frac{2 \times 0 \times 0}{0 + 0} = 0 \quad (33)$$

Updating the table to reflect the new calculations gives the results shown in Table XXIII. As the table shows, Reuters remains the best system, with Synonyms in second and Random last.

TABLE XXIII. PRECISION, RECALL, AND HARMONIC MEAN VALUES FOR DATA CLEANSING

	Reuters	Synonyms	Random
Precision	**0.3682**	0.0091	0
Recall	**0.3682**	0.0091	0
Harmonic Mean	**0.3682**	0.0091	0

B. Statistical Significance

To ensure that the results remained statistically significant once the Data Cleansing was completed, the one-way ANOVA process ran on this new set of data (the 176 submissions from the 5+keyphrase articles – but with the Random results cleansed) and the results shown in Table XXIV.

TABLE XXIV. ANOVA TABLE FOR DATA CLEANSING

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-Statistic	P-Value
Between Samples	289.7992	2	144.8996	144.2782	1.16×10^{-50}
Within Samples	527.2614	525	1.0043		
Totals	817.0606	527	1.5504		

Once again, the P-value for the likelihood of the results arising due to chance was much smaller than required to prove the results statistically significant, even smaller than the results from the first test. As the P-Value shows, the results have overwhelming statistically significance (indeed approaching 100%) – which is to say, they show strong evidence of not having arisen due to chance.

C. Discussion

The analysis of the cleansed data shows that the Reuters keyphrases remain the best system (324 of the total keyphrases, 75.7%), but the Synonyms moved up to second best (81 keyphrases, 18.9%), and Random became the weakest (23 keyphrases, 5.4%). These values are plotted in Figure 10, again on a $\log_{10}$ scale to remain consistent with Figure 7 and Figure 8.

XII. LIMITATIONS OF THE STUDIES

Outlined in this section are the limitations of the study, as well as the areas not addressed due to time and/or space issues.

The matching criteria as considered at a corpora level – either a yes or no for each paper/article/item/etc – rather than considering the number of matches of keyphrases within each paper. However, the matching does not require every author keyword/keyphrase to have a corresponding match with a keyphrase chosen by the algorithms. If all keyphrases were expected to match, then it would be expected that the number of matches recorded would be much lower than seen with the current matching criteria.

Similarly, a more detailed matching criterion could have been employed. An example of this is given by Schutz [24] and involves ranking the matches depending on the boundary conditions of the match. An exact match of a keyphrase to the 'gold standard' is scored as 1.0, whereas a sub-string might only score 0.9 (a suffix match) or 0.5 (a prefix match), and a super-string match can score 0.8 (suffix) or 0.7 (prefix). This allows a finer grained knowledge of how well the system produced keyphrases line up with the comparison keyphrases. In part this method was not employed due to maintaining consistency with earlier work on the subject.

The approaches taken in this paper mainly revolve around the 'shallow' analysis of the documents involved, rather than the 'deep' or semantic analysis. The C-Value and NC-Value include parts of this deeper analysis, as they discuss words found in 'context' to the keyphrases, but a full investigation of the semantic features of the documents was deemed outside the scope of this paper.

PubMed [12] uses the Medical Subject Headings (MeSH®) controlled vocabulary to supplement searches; however, many of the papers indexed by PubMed do not have assigned MeSH classifications, so consideration of MeSH is outside the scope of this paper – however, it is considered elsewhere in the literature [26].

XIII. DISCUSSION

The results of these two studies showed that the Synonym keyphrases were not the best in either case. In the first test, the Random keyphrases had a greater value for mean matches, but in the second test, it was the original Reuters' keyphrases. The Synonyms came second in the first test, and third in the second test. Yet, it was the original Reuters' keyphrases that came out first in the Precision, Recall, and Harmonic Mean measures for both the first and second test. Both of the tests were also determined to be statistically significant.

However, the examination of the individual results showed that the selected Random keyphrases were far from useful. Selections of keyphrases included such examples as "Which Ended" or "Total Crushing Capacity For". Therefore, the data underwent cleaning to remove the obviously suspect choices, and the calculations were repeated.

Table XXV, Table XXVI, and Table XXVII summarise the different results from this paper for each source of keyphrases. The entries in bold is the highest for that row – so it can be quickly determined from Table XXV that the Reuters keyphrases perform substantially better when only using articles that had them in abundance – as was posited as the reason for the second test in Section IX. As was seen in Table VIII, only 52.40% had keyphrases at all (11,275 articles out of 21,578), around 44% had only one keyphrase (9,443 articles, 43.76%) and less than a percent had 5 or more (125 articles, 0.58%) – the average number of keyphrases was 0.65 (14,058 keyphrases over 21,578 articles).

However, the Synonyms fared better when they did not have the Reuters keyphrases working in opposition – shown by the bold data in Table XXVI. The same can be said for the Random keyphrases in Table XXVII. From this, it can be concluded that the professional news articles writers at Reuters are good at choosing keyphrases that match the subject of the article, but that they more often than not do not assign any.

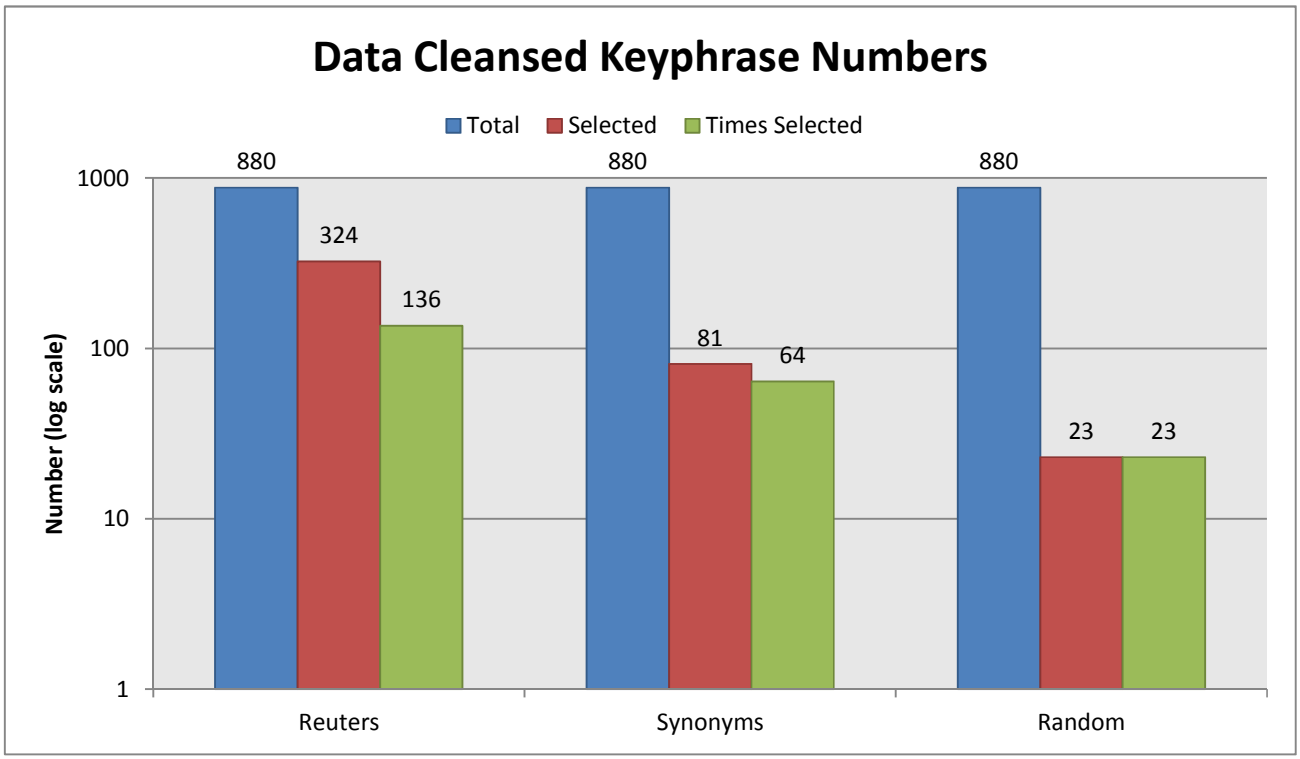


Figure 10. Graph of Data Cleansed Keyphrase Numbers (Log scale)

TABLE XXV. ALL REUTERS RESULTS

Method	First Test	Second Test	Data Cleansed
Submissions	36	**136**	**136**
Selected Keyphrases	42	**324**	**324**
Possible Selections	164	880	880
Total Keyphrases	14,058	625	625
Mean Selected	0.16800	**1.84091**	**1.84091**
Precision	0.2561	**0.3682**	**0.3682**
Recall	0.2561	**0.3682**	**0.3682**
Harmonic Mean	0.2561	**0.3682**	**0.3682**
Significance P-Value	$1.75x10^{-14}$	$3.78x10^{-33}$	$1.16x10^{-50}$

TABLE XXVI. ALL SYNONYM RESULTS

Method	First Test	Second Test	Data Cleansed
Submissions	**106**	64	64
Selected Keyphrases	**146**	81	81
Possible Selections	1250	880	880
Total Keyphrases	107,225	625	625
Mean Selected	**0.58400**	0.46023	0.46023
Precision	0.0040	**0.0091**	**0.0091**
Recall	**0.0305**	0.0091	0.0091
Harmonic Mean	**0.0305**	0.0091	0.0091
Significance P-Value	$1.75x10^{-14}$	$3.78x10^{-33}$	$1.16x10^{-50}$

The examination of the individual results showed that the selected Random keyphrases were far from useful. Selections of keyphrases included such examples as "Which Ended" or "Total Crushing Capacity For". This shows that in this instance of using human evaluators, they performed poorly – choosing inappropriate keyphrases. This may be due to the casual nature of the selection of the judges, their lack of domain specific expertise, or a failure to describe correctly the task to them via the website, or in any prior information that accompanied the distributed link.

TABLE XXVII. ALL RANDOM RESULTS

Method	First Test	Second Test	Data Cleansed
Submissions	**109**	60	23
Selected Keyphrases	**178**	95	23
Possible Selections	1250	880	880
Total Keyphrases	107,885	625	625
Mean Selected	**0.71200**	0.53977	0.13068
Precision	0.0000	0.0000	0.0000
Recall	0.0000	0.0000	0.0000
Harmonic Mean	0.0000	0.0000	0.0000
Significance P-Value	$1.75x10^{-14}$	$3.76x10^{-33}$	$1.16x10^{-50}$

However, it is also possible that, in line with the results of Sood et al. [5], humans are simply not good at choosing the correct keyphrases to assign to documents and that this is born out in the results seen in this experiment. In the automated results earlier in the paper the Random algorithm never scored more than a 14% match, yet in these results it scores up to a 71% match in the first test – over five times larger. Such a large disparity in results seems unlikely to stem from anything other than a measuring fault.

XIV. Conclusions and Further Work

Overall, therefore, the conclusions of this study are that with the data presented the original Reuters' article authors supplied the best matching keyphrases, but that the method of evaluating the results with human input was fraught with errors and likely to obfuscate the true relative worth of the algorithms. Future work will be required to examine the issues found with this method of human evaluation – and to design better test for future studies.

The results laid out earlier in this paper show that the Reuter's keyphrases are applied accurately and professionally by the writers of the news articles. Their performance ranked them as the best system in all nine of the measures (Precision, Recall, and Harmonic Mean repeated over three tests), and provided proof of their competency for the task. However, they are, by design, only of use to the articles to which they are assigned. Therefore, other than acting as a 'gold standard' to compare against, they offer no other practical use in the field of AKE.

If the Reuter's results are removed from consideration, for the above reasons, the Synonym algorithm becomes the best performing method – similarly to the Reuter's results it outperforms the Random algorithm on all nine of the measures. While this would appear to put it in a strong position of being ahead of the Random keyphrases – the actual difference in their measures is quite small. The Random algorithm, as alluded above, performs better than the other two on only one measure: Recall for the 'all articles' test (see Section VIII.A).

Acknowledgment

The authors would like to thank the School of Systems Engineering for the studentship, which enabled this project, and the contributions from the reviewers of this paper.

References

[1] R. Hussey, S. Williams, and R. Mitchell. 2012. "Automatic Keyphrase Extraction: A Comparison of Methods", Proceedings of eKNOW, The Fourth International Conference on Information, Process, and Knowledge Management, pp. 18-23. Valencia, Spain. http://www.thinkmind.org/index.php?view=article&articleid=eknow_2012_1_40_60072 [Last access: 10 December 2012]

[2] R. Hussey, S. Williams, and R. Mitchell. 2011. "A Comparison of Methods for Automatic Document Classification", Presentation at BAAL, The Forty-Fourth Annual Meeting of the British Association for Applied Linguistics. Bristol, United Kingdom.

[3] K. Frantziy, S. Ananiadou, and H. Mimaz. 2000. "Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method", International Journal on Digital Libraries, 3 (2), pp. 117-132.

[4] R. Hussey, S. Williams, and R. Mitchell. 2011. "Keyphrase Extraction by Synonym Analysis of *n*-grams for E-Journal Classification", Proceedings of eKNOW, The Third International Conference on Information, Process, and Knowledge Management, pp. 83-86. Gosier, Guadeloupe/France. http://www.thinkmind.org/index.php?view=article&articleid=eknow_2011_4_30_60053 [Last access: 10 December 2012]

[5] S.C. Sood, S.H. Owsley, K.J. Hammond, and L. Birnbaum. 2007. "TagAssist: Automatic Tag Suggestion for Blog Posts", Northwestern University. Evanston, IL, USA. http://www.icwsm.org/papers/2--Sood-Owsley-Hammond-Birnbaum.pdf [Last accessed: 10 December 2012]

[6] Technorati. 2006. "Technorati". http://www.technorati.com [Last accessed: 10 December 2012]

[7] E. Frank, G.W. Paynter, I.H. Witten, C. Gutwin, and C.G. Nevill-Manning. 1999. "Domain-Specific Keyphrase Extraction", Proceedings 16th International Joint Conference on Artificial Intelligence, pp. 668–673. San Francisco, CA Morgan Kaufmann Publishers.

[8] R. Hussey, S. Williams, R. Mitchell. 2011. "Automated Categorisation of E-Journals by Synonym Analysis of *n*-grams", International Journal on Advances in Software. Volume: 4, Number: 3 & 4, pp. 532-542. http://www.thinkmind.org/index.php?view=article&articleid=soft_v4_n34_2011_25 [Last accessed: 10 December 2012]

[9] P.M. Roget. 1911. "Roget's Thesaurus of English Words and Phrases (Index)". http://www.gutenberg.org/etext/10681 [Last accessed: 10 December 2012]

[10] M.F. Porter. 1980. "An algorithm for suffix stripping", Program, 14(3) pp. 130–137.

[11] Academics Conferences International. 2009. "ACI E-Journals". http://academic-conferences.org/ejournals.htm [Last accessed: 10 December 2012]

[12] PubMed Central. 2011. "PubMed Central Open Access Subset". http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/ [Last accessed: 10 December 2012]

[13] Reuters. 1987. "Reuters-21578 Text Categorisation Collection". http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html [Last accessed: 10 December 2012]

[14] D. Maynard and S. Ananiadou. 2000. "TRUCKS: a model for automatic multi-word term recognition", Journal of Natural Language Processing, 8 (1), pp. 101-125.

[15] Y. Matsuo and M. Ishizuka. 2003. "Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information".

[16] Y. Ohsawa, N.E. Benson, M. Yachida. 1998. "Key-Graph; Automatic indexing by co-occurrence graph based on building construction metaphor", Proceedings of the Advanced Digital Library Conference.

[17] P.D. Turney. 1999. "Learning Algorithms for Keyphrase Extraction", INRT, (pp. 34-99). Ontario.

[18] K. Barker and N. Cornacchia. 2000. "Using Noun Phrase Heads to Extract Document Keyphrases", AI '00: Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence (pp. 40-52). London: Springer.

[19] J. Carletta. 1996. "Assessing Agreement on Classification Tasks: The Kappa Statistic", Computational Linguistics, 22 (2), pp. 249-254.

[20] J. Goldstein, M. Kantrowtiz, V. Mittal, and J. Carbonell. 1999. "Summarising Text Documents: Sentence Selection

and Evaluation Metrics", Proceedings of SIGIR'99, the 22nd International Conference on Research and Development in Information Reterival, pp. 121-128. Berkeley, CA ACM Press.

[21] S. Jones, S. Lundy, and G. W. Paynter. 2002. "Interactive Document Summarisation Using Automatically Extracted Keyphrases", Proceedings of the 35th Hawaii International Conference on System Sciences 4, pp. 101-111. Hawaii IEE Computer Soceity.

[22] A. Joshi and R. Motwani. 2006. "Keyword Generation for Search Engine Advertising", IEEE International Conference on Data Mining.

[23] C. Y. Lin and E. Hovy. 2000. "The Automated Acquisiton of Topic Signatures for Text Summarisation", University of Southern California, Information Science Institute, Marina del Rey, CA.

[24] A. T. Schutz. 2008. "Keyphrase Extraction from Single Documents in the Open Domain Exploiting Linguistic and Statistical Methods", M. App. Sc Thesis.

[25] ANOVA (MathWorks – R2012a documentation). http://www.mathworks.co.uk/help/toolbox/stats/bqttcvf.html [Last accessed: 10 December 2012]

[26] L. S. Murphy, S. Reinsch, W. I. Najm, V. M. Dickerson, M. A. Seffinger, A. Adams, and S. I. Mishra. 2003. "Searching biomedical databases on complementary medicine: the use of controlled vocabulary among authors, indexers and investigators", BMC Complementary and Alternative Medicine 2003, 3:3. http://www.biomedcentral.com/1472-6882/3/3 [Last accessed: 10 December 2012]

Evaluation of the Prediction of Gene Knockout Effects by Minimal Pathway Enumeration

Takehide Soh†
Kobe University
1-1, Rokko-dai, Nada, Kobe
Hyogo, Japan
soh@lion.kobe-u.ac.jp

Katsumi Inoue
National Institute of Informatics
2-1-2, Hitotsubashi, Chiyoda-ku,
Tokyo, Japan
inoue@nii.ac.jp

Tomoya Baba
Transdisciplinary Research
Integration Center
1111 Yata, Mishima,
411-8540, Japan
tobaba@lab.nig.ac.jp

Toyoyuki Takada, Toshihiko Shiroishi
National Institute of Genetics
1111 Yata, Mishima,
411-8540, Japan
{ttakada,tshirois}@lab.nig.ac.jp

***Abstract*—In this paper, we propose a method to predict gene knockout effects for the cell growth by utilizing biological databases such as KEGG and EcoCyc, in which biological knowledge and experimental results have been collected. At first, biological networks are constructed from such databases and configure experimental conditions by giving source metabolites, target metabolites, and knockout genes. All minimal active pathways are then enumerated, which are minimal subsets of a given network using source metabolites to produce target metabolites. Finally, the effects of gene knockouts are predicted by measuring the difference of minimal active pathways between original networks and knockout ones. In the experiments, we applied it to predict the single gene knockout effects on the glycolysis pathway and amino acids biosynthesis in *Escherichia coli*. We also analyze which pathways are important to *Escherichia coli* and predict lethal pairs of knockout genes. In the results, our method predicted three out of four essential genes, which agree with the biological results of the Keio collection containing comprehensive cell growth data. In addition, predicted lethal pairs of genes also agree with the biological results of double gene knockouts.**

Keywords-metabolic pathways; gene knockout; prediction method; minimal pathway; Keio collection.

I. INTRODUCTION

This paper is an extended version of the previously published conference paper [1]. While the earlier paper only considered effects of single gene knockouts, this paper proceeds the analysis of those effects and shows the prediction result of double gene knockouts, which is compared with biological results of [2].

Living organisms, such as bacteria, fishes, animals, and humans, are kept alive by a huge number of intracellular chemical reactions. In *systems biology*, interactions of such chemical reactions are represented in a network called a *pathway*. Pathways have been actively researched in the last decade [3]–[5]. In addition, it is a biologically important subject to reveal the function of genes, which affect the phenotype of organisms. For model organisms such as *Escherichia coli* (*E. coli*), it has been approached by various methods. Constructing gene knockout organisms is an example of such methods [6]–[8]. However, it generally involves high costs and is limited by target genes and organisms.

In this paper, we propose a computation method to predict gene knockout effects by identifying *active pathways*, which are sub-pathways that produce target metabolites from source metabolites. We particularly focus on *minimal active pathways* [9], which do not contain any other active pathways. In other words, all elements of each minimal active pathway are qualitatively essential to produce target metabolites. To predict gene knockout effects by the enumeration of minimal active pathways, at first, *extended pathways*, which include relations between enzymatic reactions and genes, are introduced. Then, the problem of finding minimal active pathways on the extended pathway with gene knockouts is formalized. After computing the solution of the problem, our method predicts gene knockout effects by collecting minimal active pathways that are still active under given gene knockouts.

To evaluate our method, *E. coli* is chosen as our target organism, since it has been studied and much information is available on public resources such as KEGG [10] and EcoCyc [11]. The proposed method is applied to predict gene knockout effects on *E. coli* utilizing biological databases KEGG and EcoCyc, in which biological knowledge and experimental results have been collected. In the experiments, we compare our prediction with the cell growth of every single gene knockout *E. coli* strain, which was obtained from the Keio collection [6]. In addition to the above experiments, using the biological results of [2], we analyze which minimal active pathways are important and predict lethal pairs of gene knockouts for *E. coli*. We also apply our method to predict the gene knockout effects on amino acid biosynthesis and discuss which reactions are suspected to be lacked in databases.

This paper is organized as follows. At first, databases used in this paper and our research framework are explained in Section II. The extended pathway is defined in Section III. Then, Section IV formalizes the problem of finding minimal

†This work has been done while affiliated with Transdisciplinary Research Integration Center, 2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo, Japan.

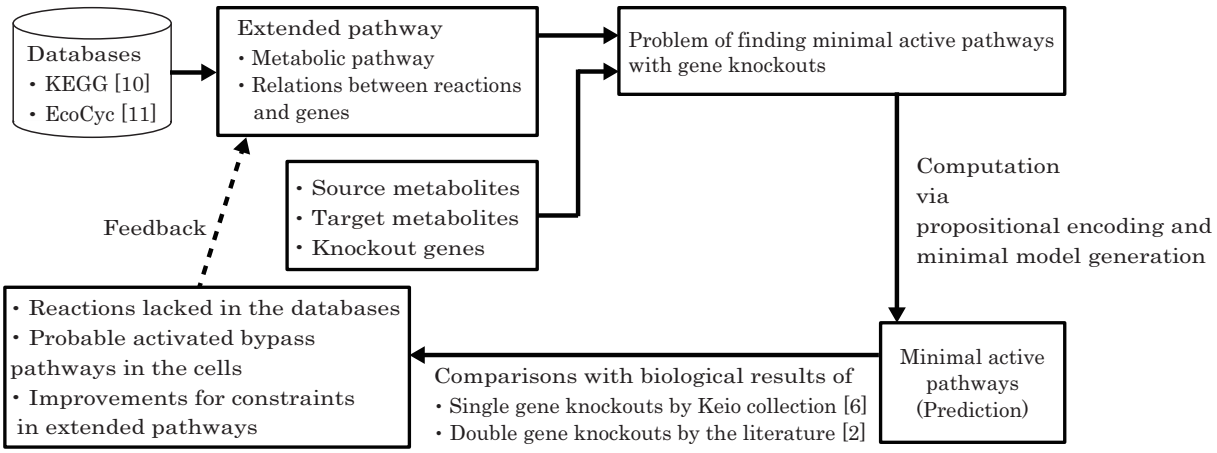


Figure 1. Used Databases and our Research Framework

active pathways on the extended pathway. The effect of gene knockouts is formalized in Section V. Following that, our computational method is shown in Section VI. In Section VII, computational predictions are compared with results of biological experiments, and we have discussions. Following Section VIII of related work, this paper is concluded in Section IX.

II. Used Databases and Research Framework

This section explains used databases and our research framework shown in Figure 1. In this paper, we particularly focus on *E. coli*. The metabolic pathway has been revealed by biochemical, molecular, and genetic studies, and *E. coli* is the organism in most detail. A large number of *E. coli* studies has contributed to several kinds of biological databases. In particular, we used the following two databases to construct our input network, called an extended pathway.

One is *EcoCyc* [11]. It is a bioinformatics database that describes the genome and the biochemical machinery of *E. coli* K-12 MG1655. The EcoCyc project performs literature-based curation of the entire genome, metabolic pathways, etc. Specifically, it has been doing a literature-based curation from more than 19,000 publications. We constructed metabolic pathways with EcoCyc. The other one is *Kyoto encyclopedia of genes and genomes* (KEGG) [10], which is a database resource that integrates genomic, chemical, and systemic functional information. In particular, gene catalogs in the completely sequenced genomes, from bacteria to humans, are linked to higher-level systemic functions of the cell, the organism, and the ecosystem. A distinguished feature of KEGG is that it provides useful application program interfaces (API). We connected enzymatic reactions of metabolic pathways to genes with this API.

Figure 1 shows our research framework using the two databases. At first, the input network called *extended pathway* is constructed from those databases. Then the problem of finding minimal active pathways is constructed by giving source metabolites, target metabolites, knockout genes, and the extended pathway. Then, we compute minimal active pathways using source metabolites to produce target metabolites. In the case of wild cells, we usually obtain multiple minimal active pathways including bypass pathways. However, in the case of knockout cells, we lose some (or all) of them. In brief, we predict the effects of gene knockouts from how many pathways are lost from the case of wild cells.

To evaluate our prediction method, we usually need additional biological experiments. However, Baba *et al.* comprehensively experimented on the cell growth of every single gene knockout strain [6]. Thanks to this research, we can evaluate our method with comparative ease. We briefly explain this research as follows. The *E. coli* K-12 single gene knockout mutant set, named *Keio collection*, is constructed as a resource for systems biological analyses. Excluding repetitive genes, e.g., insertion sequences related genes, 4288 protein coding genes are targeted for the systematic single gene knockout experiments. Of those, 3985 genes are successfully disrupted, and those of single gene knockout mutants are constructed as the Keio collection. On the other hand, 303 genes are not disrupted and they are thought to be essential gene candidates. Those single gene knockout mutants have the same genome background, which results in an advantage for distinct functional analysis of the targeted gene. The genome-wide relationship between the genome structure, i.e., genotype, and the phenomena, i.e., phenotype, which are analyzed by using the Keio collection has become available. In addition to the Keio collection, results of double gene knockouts by the literature [2] is also used.

After the above evaluation, some differences between prediction results and biological results of [6] and [2] will be found. Those differences are used to found lacked reactions, bypass pathways actually used in cells, and improvements for pathway model, which refine extended pathways.

Although Figure 1 shows specific databases for *E. coli*,

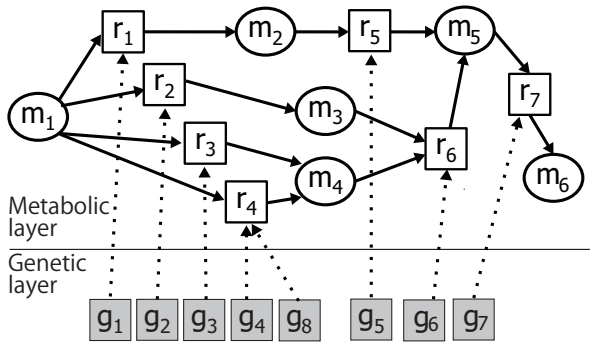


Figure 2. Example of an Extended Pathway

the research framework itself can be applied for other organisms whose pathway information is available. For instance, although there is a large difference between *E. coli* and mice, the central metabolism is similar and it could be a potential application.

III. Extended Pathways

In this section, we explain how to represent metabolic pathways and their relations to genes. We then define the extended pathway.

To represent metabolic pathways, we commonly use bipartite directed graph representation as follows. Let M be a set of metabolites and R be a set of reactions. For M and R, $M \cap R = \emptyset$ holds. Let $A_M \subseteq (R \times M) \cup (M \times R)$ be a set of arcs. A *metabolic pathway* is represented in a directed bipartite graph $\mathcal{G}_\mathcal{M} = (M \cup R, A_M)$, where M and R are two sets of nodes, and A_M is a set of arcs. In addition to the metabolic pathway, we consider relations between enzymatic reactions and genes. Let G be a set of genes and A_G be a set of arcs such that $A_G \subseteq (G \times R)$. That is, A_G represents relations between enzymatic reactions and genes. Let N be a set of nodes such that $N = M \cup R \cup G$ and A be a set of arcs such that $A = A_M \cup A_G$. Then, the *extended pathway* is represented in a directed graph $\mathcal{G} = (N, A)$.

Figure 2 shows an example of the extended pathway. As the figure shows, it consists of two layers: the metabolic layer and the genetic layer. The genetic layer is the difference between the metabolic pathway and the extended pathway. In this example, the pathway consists of nodes of $M = \{m_1, m_2, \ldots, m_6\}$, $R = \{r_1, r_2, \ldots, r_7\}$, and $G = \{g_1, g_2, \ldots, g_8\}$. Each arc represents relations between elements. For instance, the activation of the reaction r_6 needs the production of metabolites m_3 and m_4 and the expression of g_6. We will explain the interpretation of the extended pathway in detail in the next section.

IV. Minimal Active Pathways with Gene Knockouts

In this section, we explain the notion of *producible*, *activatable* and *minimal active pathway* on the extended pathway, while the minimal active pathway is introduced only on the metabolic pathway in the literature [9].

We here define $M_S \subset M$ as a set of source metabolites and $M_T \subset M$ as a set of target metabolites such that $M_S \cap M_T = \emptyset$. An *extended pathway instance* is represented in a four tuple $\pi = (N, A, M_S, M_T)$, where $N = M \cup R \cup G$, $A = A_M \cup A_G$. Let K be a set of genes such that $K \subseteq G$. We use K as a set of knockout genes in a given pathway. A *knockout instance* is represented in a five tuple $\pi_K = (N, A, M_S, M_T, K)$. If $K = \emptyset$ then π_K corresponds to π.

Let m, r be a metabolite and a reaction such that $m \in M$ and $r \in R$, respectively. A metabolite $m \in M$ is called a *reactant* of a reaction $r \in R$ when there is an arc $(m, r) \in A$. On the other hand, a metabolite $m \in M$ is called a *product* of a reaction $r \in R$ when there is an arc $(r, m) \in A$. Furthermore, a gene $g \in G$ is called a *corresponding gene* of a reaction $r \in R$ when there is an arc $(g, r) \in A$.

A reaction is called a *reversible reaction* if it can occur in both directions between reactants and products. In this paper, we distinguish a reversible reaction as two reactions. Suppose that there is a reversible reaction r_1 that has m_1 and m_2 as reactants and m_3 and m_4 as products. In this case, we split the reaction r_1 into two reactions r_{1a} and r_{1b} such that one of them has m_1 and m_2 as products and m_3 and m_4 as reactants.

Let $s : R \to 2^M$ be a mapping from a set of reactions to a power set of metabolites such that $s(r) = \{m \in M \mid (m, r) \in A\}$ represents the set of metabolites that are needed to turn the reaction r to be active. Let $p : R \to 2^M$ be a mapping from a set of reactions to a power set of metabolites such that $p(r) = \{m \in M \mid (r, m) \in A\}$ represents the set of metabolites that are produced by the reaction r. Let $c : R \to 2^G$ be a mapping from a set of reactions to a power set of genes such that $c(r) = \{g \in G \mid (g, r) \in A\}$ represents the set of genes that are corresponding genes of the reaction r. Let $p' : M \to 2^R$ be a mapping from a set of metabolites to a power set of reactions such that $p'(m) = \{r \in R \mid (r, m) \in A\}$. Let $c' : G \to 2^R$ be a mapping from a set of genes to a power set of reactions such that $c'(g) = \{r \in R \mid (g, r) \in A\}$.

Let t be an integer variable representing time. In this paper, the time is used to represent order relation between reactions to produce target metabolites from source metabolites. In the following, we explain important notions related to production of metabolites, activation of reactions, and expression of genes. Since we focus on gene knockouts, we suppose that almost all genes exist in the cell of a given organism. We also suppose that if genes exist, then they are expressed and available to construct enzymes needed for enzymatic reactions. The reason for this condition is that we want to simulate how the lack of corresponding genes affects metabolic pathway rather than how the existence of genes affects other elements. Although our pathway modeling is simple, it allows us to analyze a whole cell scale pathway.

Let $\pi_K = (N, A, M_S, M_T, K)$ be a knockout instance, where $N = M \cup R \cup G$, $A = A_M \cup A_G$. Let $\mathcal{G} = (N, A)$ be an extended pathway. Let $M' \subset M$ be a subset of metabolites. A metabolite $m \in M$ is obviously *producible* at time $t = 0$ from M' on $\mathcal{G}$ if $m \in M'$ holds. A reaction $r \in R$ is *activatable* at time $t > 0$ from M' on $\mathcal{G}$ if the following two conditions are satisfied: (i) for every $m \in s(r)$, m is producible at time $t-1$ from M', (ii) at least one corresponding gene $g \in c(r)$ is not included in K. A metabolite $m \in M$ is *producible* at time $t > 0$ from M' on $\mathcal{G}$ if there is at least one activatable reaction r at time t such that $m \in p(r)$. If r is activatable at time t, then r is activatable at time $t+1$. If m is producible at time t, then m is producible at time $t+1$.

Let $\mathcal{G}' = (N', A')$ be a sub-graph of $\mathcal{G}$, where $N' = M' \cup R' \cup G'$ and $A' = A'_M \cup A'_G$. Then, an active pathway of $\pi_K = (N, A, M_S, M_T, K)$ is defined as follows.

Definition 1: Active Pathway of Knockout Instance
A bipartite directed graph $\mathcal{G}'$ is an *active pathway* of π_K if it satisfies the following conditions:

- $M_T \subset M'$
- $M' = M_S \cup \{m \in M \mid (m, r) \subseteq A, r \in R'\} \cup \{m \in M \mid (r, m) \subseteq A, r \in R'\}$
- $A' = \{(m, r) \in A \mid r \in R'\} \cup \{(r, m) \in A \mid r \in R'\} \cup \{(g, r) \in A \mid g \notin K, r \in R'\}$
- $G' = \{g \in G \mid (g, r) \in A', r \in R'\}$
- For every $m \in M'$, m is producible from M_S on $\mathcal{G}'$

From Definition 1, active pathways include a set of metabolites, reactions, and genes, which are producible and activatable from M_S on $\mathcal{G}'$ such that all target metabolites M_T become producible. The number of active pathways depends on the combination of M_S and M_T but an extended pathway generally has a large number of active pathways. We thus particularly focus on minimal ones rather than active pathways. We give the definition of minimal active pathways of π_K as follows. Let $\mathcal{G}$ and $\mathcal{G}'$ be extended pathways. We say that $\mathcal{G}$ is *smaller* than $\mathcal{G}'$ and represented in $\mathcal{G} \subset \mathcal{G}'$ if $R \subset R'$. An active pathway $\mathcal{G}$ is *minimal active pathway* of π_K *iff* there is no active pathway of π_K, which is smaller than $\mathcal{G}$. As this definition shows, we only need to see sets of reactions to compare two pathways. Thus, in the rest of this paper, we sometimes represent a minimal active pathway as a set of reactions.

Any reactions included in a minimal active pathway cannot be deleted to produce target metabolites. Intuitively, this indicates that each of the elements of a minimal active pathway is essential. In practice, minimal active pathways including a large number of reactions are considered to be biologically inefficient. We thus introduce a time limitation z and pathways that can make all target metabolites producible by $t = z$. In the following, we consider the problem of finding minimal active pathways with respect to π_K and z.

V. KNOCKOUT EFFECTS

This section provides how to predict knockout effects. In the following, we give some definitions for the prediction. Let $\pi = (N, A, M_S, M_T)$ and $\pi_K = (N, A, M_S, M_T, K)$ be an extended pathway instance and a knockout instance, respectively. In addition, we denote the number of minimal active pathways of π as $|\pi|$ and the number of minimal active pathways of π_K as $|\pi_K|$. Obviously, $|\pi_K| \leq |\pi|$ holds. Then, the gene knockout effect, i.e., the prediction by the proposed method, is given by $E_K = |\pi| - |\pi_K|$. Let K_a and K_b be sets of knockout genes. If $E_{K_a} > E_{K_b}$ holds, then we say that the gene knockout effect of K_a is stronger than that of K_b. If $|\pi_K| = 0$, i.e., $E_K = |\pi|$, then we say that the knockout effect of K is *critical* to produce target metabolites. Various metabolites are known as vital metabolites, which means organisms cannot survive without them. That is, if some gene knockouts are critical to produce such metabolites, then a given organism cannot grow any more or dies. If $|K| = 1$ and its effect is critical to produce vital metabolites, then we say that the gene $g \in K$ is *essential*.

In the following, we explain the above definition with a specific example. Suppose that we are given a pathway instance $\pi = (N, A, M_S, M_T)$, where N and A are from the extended pathway shown in Figure 2, and the source metabolite is $M_S = \{m_1\}$ and the target metabolite is $M_T = \{m_6\}$. Obviously, $|\pi| = 3$ and the minimal active pathways of π are specifically as follows: $\{r_1, r_5, r_7\}, \{r_2, r_3, r_6, r_7\}, \{r_2, r_4, r_6, r_7\}$. Then, we consider the following knockout instances π_{K_1} and π_{K_2}, where $K_1 = \{g_1\}$ and $K_2 = \{g_6\}$. For π_{K_1}, minimal active pathways including r_1 can no longer be solutions, i.e., $|\pi_{K_1}| = 2$. For π_{K_2}, minimal active pathways including r_6 can no longer be solutions either. Thus, $\{r_2, r_3, r_6, r_7\}$ and $\{r_2, r_4, r_6, r_7\}$ are deleted from the solutions of π, i.e., $|\pi_{K_2}| = 1$. Consequently, we can say that the knockout effect of K_2 is stronger than that of K_1. Moreover, suppose that $K_3 = \{g_7\}$. Then, there is no minimal active pathway of π_{K_3} and we say that the knockout effect of K_3 is critical to produce m_6. If m_6 is a vital metabolite, we can simultaneously say that g_7 is an essential gene.

In addition to the number of remaining minimal active pathways after knockouts, an important factor in the prediction is the gain of ATPs. This is because pathways that are inefficient with respect to energy consumption will not be used in organisms. Let $|\pi^{a+}|$, $|\pi_K^{a+}|$ be the number of minimal active pathways of π and π_K, which gain ATPs, respectively. Then, the gene knockout effect with respect to ATP production is given by $E_K^{a+} = |\pi^{a+}| - |\pi_K^{a+}|$. In particular, it is important when we consider the glycolysis pathway since one of its main functions is to gain ATPs. However, we cannot find any pathways producing ATPs on some other pathways, i.e., minimal active pathways on them must consume ATPs. In this case, the number of minimal

active pathways, which consume fewer ATPs, should be considered instead of $|\pi^{a+}|$ and $|\pi_K^{a+}|$.

VI. Computational Method

This section explains how to compute $|\pi_K|$. In this paper, we use the method of computing all minimal active pathways of π by Soh and Inoue [9]. This method computes pathways through propositional encoding and minimal model generation. An advantage is that this method is flexible for adding biological constraints, which is explored in [9]. Moreover, we can utilize SAT technologies, which have been developed actively in recent years.

In the following, we briefly explain the propositional encoding to compute minimal active pathways of π. Let i, j be integers denoting indices for metabolites and reactions. Let t be an integer variable representing time. Let $\pi = (N, A, M_S, M_T)$ be an extended pathway instance, where $N = M \cup R \cup G$, $A = A_M \cup A_G$. We introduce two kinds of propositional variables. Let $m_{i,t}^*$ be a propositional variable, which is $true$ if a metabolite $m_i \in M$ is producible at time t. Let $r_{j,t}^*$ be a propositional variable, which is $true$ if a reaction $r_j \in R$ is activatable at time t.

The encoding of the problem of finding minimal active pathways with respect to π_K and z is as follows.

$$\psi_1 = \bigwedge_{0 \le t < z} \bigwedge_{m_i \in M} \left(m_{i,t}^* \rightarrow m_{i,t+1}^*\right)$$

$$\psi_2 = \bigwedge_{0 \le t < z} \bigwedge_{r_j \in R} \left(r_{j,t}^* \rightarrow r_{j,t+1}^*\right)$$

$$\psi_3 = \bigwedge_{1 \le t \le z} \bigwedge_{r_j \in R} \left(r_{j,t}^* \rightarrow \bigwedge_{m_i \in s(r_j)} m_{i,t-1}^* \right)$$

$$\psi_4 = \bigwedge_{1 \le t \le z} \bigwedge_{r_j \in R} \left(r_{j,t}^* \rightarrow \bigwedge_{m_i \in p(r_j)} m_{i,t}^* \right)$$

$$\psi_5 = \bigwedge_{m_i \in (M \setminus M_S)} \bigwedge_{1 \le t \le z} \left(m_{i,t}^* \rightarrow m_{i,t-1}^* \vee \bigvee_{r_j \in p'(m_i)} r_{j,t}^* \right)$$

$$\psi_6 = \bigwedge_{m_i \in M_S} m_{i,0}^* \wedge \bigwedge_{m_{i'} \in M \setminus M_S} \neg m_{i',0}^*$$

$$\psi_7 = \bigwedge_{m_i \in M_T} m_{i,z}^*$$

The formulas ψ_1 and ψ_2 represent that once a metabolite (or a reaction) is made to producible (or activatable), then it remains in the producible (or activatable) state. The formula ψ_3 represents that if a reaction r_j is activatable at time t then its reactants must be producible at time $t-1$. The formula ψ_4 represents that if a reaction r_j is activatable at time t then its products must be producible at time t. The formula ψ_5 represents that if a reaction m_i is producible then either two states hold: the metabolite m_i is producible at $t-1$ or at least one reaction r_j is activatable. The formulas ψ_6 and ψ_7 represent source metabolites and target metabolites. We denote the conjunction of $\psi_1, \ldots, \psi_7$ as Ψ_z. Then, we can enumerate minimal active pathways with respect to π_K and z by computing minimal models of Ψ_z with respect to $V^z = \{r_{i,z}^* | r_i \in R\}$.

The computation for π is always needed to compare a wild cell and its mutant. We thus explain a method to compute all minimal active pathways of π_K for a set of knockout genes K. Actually, when the minimal active pathways of π are obtained, we do not need much additional computation. All minimal active pathways of π_K are obtained by selecting pathways that do not contain some $r \in R_K$, where $R_K = \{r \in c'(g) \mid g \in K\}$. The procedure is given as follows: (i) enumerate all minimal active pathways with respect to π and z, (ii) delete minimal active pathways including some $r \in R_K$, where $R_K = \{r \in c'(g) \mid g \in K\}$. As well as the above procedure, there is another way to compute all minimal active pathways with respect to π_K and z. The same is achieved by adding constraints, which inhibit the activation of each reaction in R_K, to the formula Ψ_z.

VII. Experimental Results

This section provides experimental results and discussions. At first, we describe experimental conditions. Then, we show the results of our prediction of knockout effects for glycolysis and amino acids biosynthesis.

A. Experimental conditions

We constructed extended pathways from EcoCyc [11] and KEGG [10]. Specifically, we use EcoCyc to construct metabolic pathways, which consists of 1222 metabolites and 1920 reactions. Moreover, we use KEGG to construct relations between enzymatic reactions and genes. The entire extended pathway we used is constructed from these two databases. In the following experiments, we denote a reversible reaction in EcoCyc as two differently directed two reactions by adding suffixes _a and _b, respectively. In addition, some reactions such as *6PGLUCONDEHYDROG-RXN* can accept different metabolites as its input, e.g., *6PGLUCONDEHYDROG-RXN* is considered to be able to use NAD^+ and $NADP^+$. In this case, we distinguish it as two different reactions by adding suffixes _1 and _2, respectively.

Each experiment has been done using a PC (3.2GHz CPU) running on OS X 10.6. For computation, we use a SAT solver Minisat2 [12]. Koshimura *et al.* proposed a procedure computing minimal models with SAT solvers [13]. We follow their procedure to generate minimal models by using a SAT solver.

To evaluate our method, we use the Keio collection as is described in Section II. In particular, we use their results on the MOPS medium whose main nutrient is glucose.

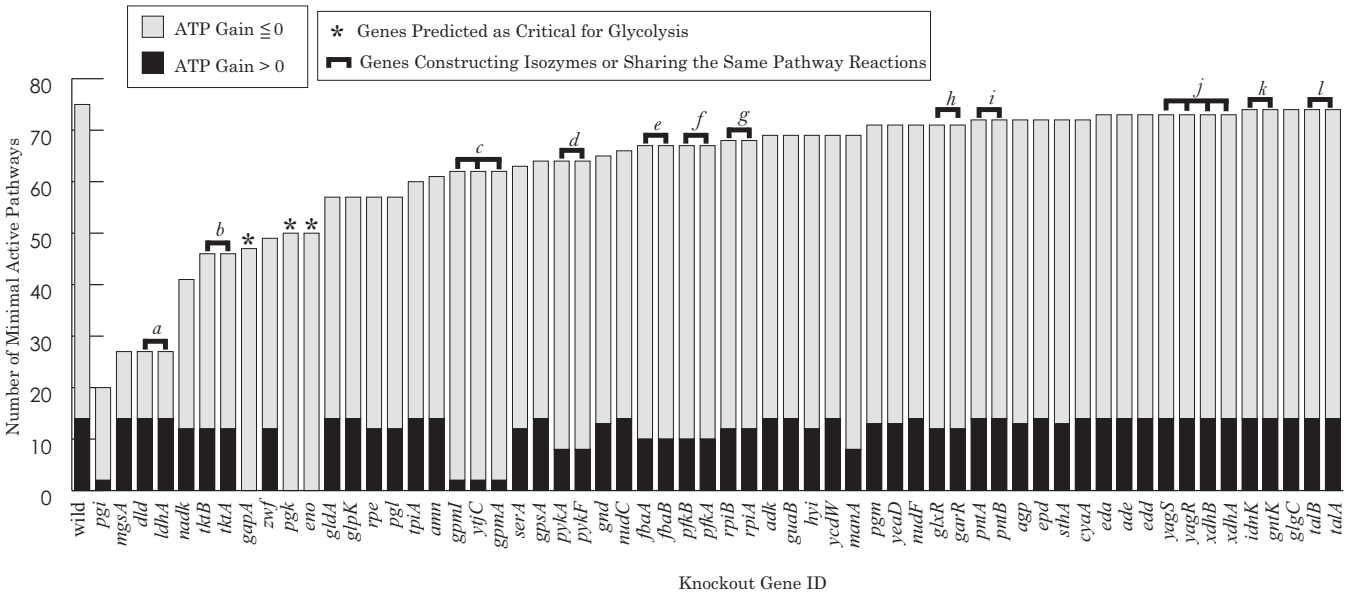
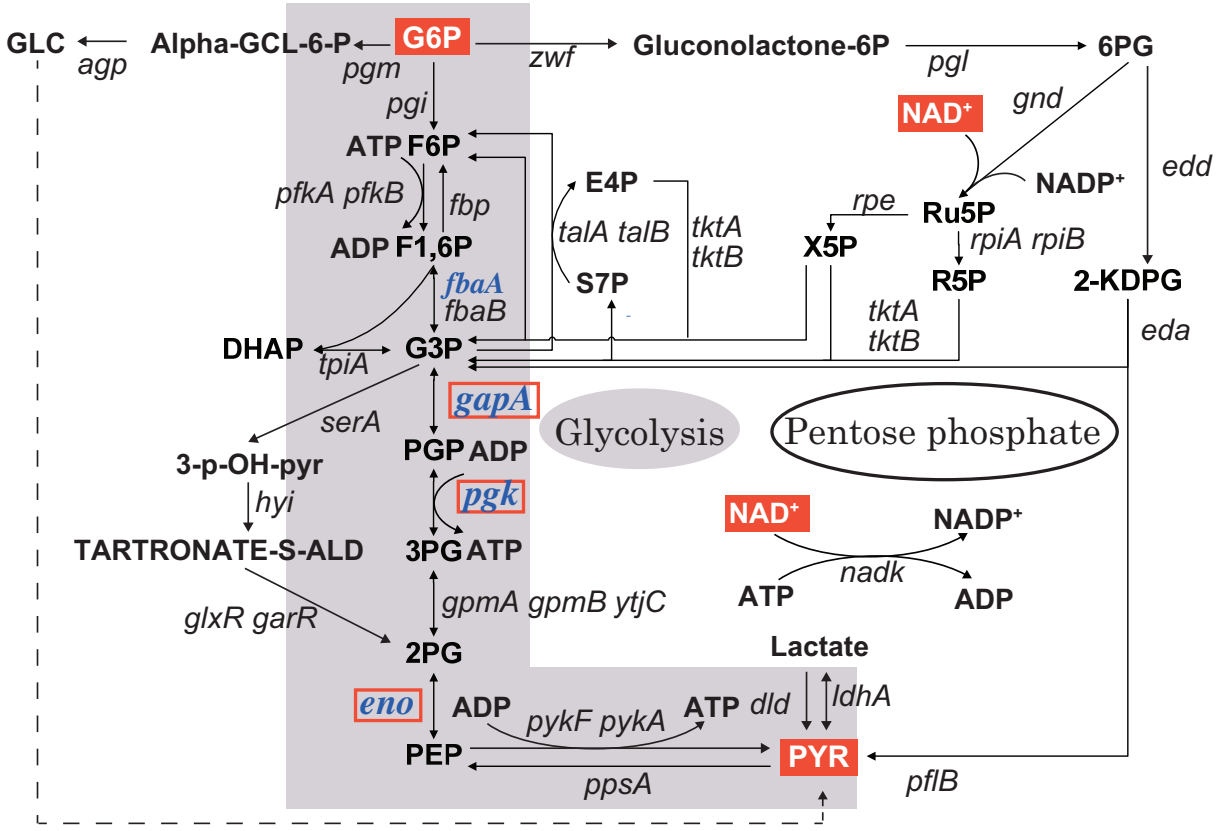


Figure 3. The Number of Minimal Active Pathways for each Gene Knockout on Glycolysis

Table I
SINGLE GENE KNOCKOUTS FOR GLYCOLYSIS
Symbols "*" denote genes predicted as critical and "a" to "l" denote genes constructing isozymes or sharing the same pathway reactions.

Gene	#Minimal Active Pathways			Keio Collection [6]	
	Total	ATP Gain		MOPS24hr	MOPS48hr
		plus	others		
wild	75	14	61	0.219-0.392	0.216-0.480
pgi	20	2	18	0.137	0.542
mgsA	27	14	13	0.293	0.371
dld^a	27	14	13	0.303	0.366
ldhA^a	27	14	13	0.357	0.393
nadk (yfjB)	41	12	29	N.A.	N.A.
tktB^b	46	12	34	0.311	0.315
tktA^b	46	12	34	0.317	0.327
*gapA**	47	0	47	N.A.	N.A.
zwf	49	12	37	0.231	0.223
*pgk**	50	0	50	N.A.	N.A.
*eno**	50	0	50	N.A.	N.A.
gldA	57	14	43	0.255	0.351
glpK	57	14	43	0.283	0.409
rpe	57	14	43	0.347	0.335
pgl(ybhE)	57	12	45	0.551	0.686
tpiA	60	14	46	0.345	0.321
amn	61	14	47	0.330	0.342
gpmI^c	62	2	60	0.303	0.292
ytjC^c	62	2	60	0.339	0.378
gpmA^c	62	2	60	0.383	0.240
serA	63	12	51	0.007	0.021
gpsA	64	14	50	N.A.	N.A.
pykA^d	64	8	56	0.266	0.299
pykF^d	64	8	56	0.310	0.320
gnd	64	12	52	0.251	0.282
nudC	66	14	52	0.445	0.518
fbaA^e	67	10	57	N.A.	N.A.
fbaB^e	67	10	57	0.371	0.447
pfkB^f	67	10	57	0.270	0.258
pfkA^f	67	10	57	0.087	0.554

Gene	#Minimal Active Pathways			Keio Collection [6]	
	Total	ATP Gain		MOPS24hr	MOPS48hr
		plus	minus		
rpiB^g	68	12	56	0.326	0.394
rpiA^g	68	12	56	0.340	0.372
adk	69	14	55	N.A.	N.A.
guaB	69	14	55	0.005	0.020
hyi	69	12	57	0.191	0.197
ycdW	69	14	55	0.255	0.301
manA	69	8	61	0.334	0.355
pgm	71	13	58	0.169	0.158
yeaD	71	13	58	0.233	0.289
nudF	71	14	57	0.376	0.373
glxR^h	71	12	59	0.226	0.231
garR^h	71	12	59	0.400	0.368
pntA^i	72	14	58	0.220	0.288
pntB^i	72	14	58	0.317	0.513
agp	72	13	59	0.319	0.528
epd	72	14	58	0.321	0.344
sthA	72	14	58	0.271	0.515
cyaA	72	14	58	0.455	0.295
eda	73	14	59	0.211	0.205
ade	73	14	59	0.246	0.304
edd	73	14	59	0.269	0.282
yagS^j	73	14	59	0.273	0.125
yagR^j	73	14	59	0.295	0.306
xdhB^j	73	14	59	0.324	0.447
xdhA^j	73	14	59	0.433	0.488
idnK^k	74	14	60	0.230	0.140
gntK^k	74	14	60	0.303	0.303
glgC	74	14	60	0.304	0.327
talB^l	74	14	60	0.254	0.169
talA^l	74	14	60	0.316	0.345



Source metabolites (M_S): β-D-glucose-6-phosphate, NAD+	G6P NAD+
Target metabolite (M_T): pyruvate	PYR
Four essential genes confirmed by KEIO collection	*fbaA* *gapA* *pgk* *eno*
Genes predicted by our method	

Figure 4. Glycolysis and Pentose Phosphate Pathways of *E. coli* from the literature by Ishii *et al.* [8]

In addition, we use the results of the literature [2] for a comparison for double gene knockouts. For the former results, i.e., Keio collection, we consider that the set of knockout genes K consists of one gene. The cell growth of the wild cell is ranged from 0.216 to 0.480 in Keio collection and 0.378 in the literature [2]. For both data, if cell growth is less than 0.1, which is less than half of them, we then say that the cell is strongly affected by a gene knockout.

B. Results for Glycolysis Analysis

First, we analyze the glycolysis pathway of *E. coli*. In accordance with the MOPS medium of the Keio collection [6], a set of source metabolites M_S is chosen as follows: {β-D-glucose-6-phosphate, H^+, H_2O, ATP, ADP, phosphate, and NAD^+}. In addition, pyruvate is given as the target metabolite to analyze glycolysis, i.e., $M_T = \{\text{pyruvate}\}$. We then compute all minimal active pathways from the entire metabolic pathway of *E. coli*. As we can see in biological literature such as the work of Ferguson *et al.* [14], glycolysis is known to a pathway constructed by eight steps. However, if some reactions are disabled, then *E. coli* is expected to use other bypass pathways by using additional reactions. In this experiment, we consider four additional reactions, i.e., the number of reactions included in each pathway is limited to less than or equal to 12 as well as $z = 12$.

At first, we computed all minimal active pathways with the above conditions and obtained 75 minimal active pathways from the entire reactions database, which consists of 1920 reactions. We then connect 61 genes to reactions by API on KEGG. Since there is no data, some reactions are remaining unconnected. Next, we computed minimal active pathways with each gene knockout. This experiment was done within four seconds. Figure 3 shows the results of 61

gene knockouts. The x-axis denotes each gene knockout and the y-axis denotes the number of minimal active pathways. As is shown in the figure, we compute minimal active pathways of $\pi_{K_1}, \ldots, \pi_{K_{61}}$ such that $K_1 = \{pgi\}, K_2 = \{mgsA\}, \ldots, K_{61} = \{talA\}$. However, since some of the 61 genes construct isozymes, such single gene knockout K_i does not affect the number of minimal active pathways $|\pi_{K_i}|$. However, for reference, we compute the effect of the gene knockouts that disables all of isozymes. For instance, *tktA* and *tktB* construct isozymes. In this case, the number of minimal active pathways in the figure shows the case of the gene knockout of both *tktA* and *tktB*. For each gene knockout, we computed the gain of ATP in each minimal active pathway, which is calculated by counting the number of both reactions with the coefficient of ATP: ones consuming ATP and the other ones producing ATP. Minimal active pathways that produce the positive number of ATPs are more important than the others because producing ATP is a main function of glycolysis.

From the figure, we can see that *E. coli* keeps almost all minimal active pathways even by more than half of single gene knockouts. This is considered to indicate the robustness of *E. coli*. However, some gene knockouts dramatically reduce the number of minimal active pathways. In particular, the single gene knockouts of *gapA*, *pgk*, and *eno* destroy all minimal active pathways producing ATPs. Thus, they are predicted to strongly affect the glycolysis of *E. coli*.

To evaluate the above predictions, we compare them with the Keio collection. Table I compares all gene knockouts shown in Figure 3 regarding the number of lost minimal active pathways. Column 1, Gene, shows gene names except *wild*, which denotes an empty set of knockout genes, i.e., $K = \emptyset$. Other rows denote the result of single gene knockout. Column 2, Total, shows the total number of minimal active pathways, i.e., $|\pi_{K_i}|$. Columns 3 and 4 show the number of minimal active pathways, which gain ATPs, i.e, $|\pi_{K_i}| > 0$ and consume ATPs, i.e., $|\pi_{K_i}| \leq 0$. Column 7, MOPS24hr, and Column 8, MOPS48hr, show the cell growth of *E. coli* after 24 hours and 48 hours, respectively. Note that N.A. (not applicable) refers to essential genes [6]. As the first row of Table I shows, we found 14 minimal active pathways that produce the positive number of ATPs on the wild cell of *E. coli* while there are 75 in total[1].

Distinguished single gene knockouts are $K_8 = gapA$, $K_{10} = pgk$, and $K_{11} = eno$. Each gene knockout effect with respect to ATP production is $E^{a+}_{K_8} = E^{a+}_{K_{10}} = E^{a+}_{K_{11}} = 14$ and it is the strongest gene knockout effect with respect to ATP production, which is the important function of glycolysis. For this prediction, the Keio collection shows "N.A." for each gene knockout. Thus, in glycolysis, our predictions successfully agree with the results of the Keio collection. However, there are other gene knockouts showing "N.A." in the results of Keio collection, that is, *gpsA*, *fbaA*, *nadk (yfjB)* and *adk*. For those genes, the number of minimal active pathways is not so reduced. Then, in those gene knockouts, it can be considered that *E. coli* is damaged in other pathways rather than the glycolysis pathway, discussing at the following sections.

[1]Those 75 minimal active pathways are shown in a supporting online material in http://kix.istc.kobe-u.ac.jp/~soh/supplement/prediction.html.

C. Discussion for Glycolysis Analysis

In this section, we first discuss about the difference of our prediction and the cell growth of the Keio collection. Figure 4 shows the glycolysis pathway modified from the one in the literature [8]. We pick up the figure of glycolysis and pentose phosphate pathways. Abbreviation is same as the literature [8]. Each node, e.g., G6P, denotes a metabolite and edge denote chemical reactions. Labels of edges denote corresponding genes to reactions. A dotted line denotes the abstraction of some reactions whose genes are not registered in the databases. The figure also shows generally known four essential genes in terms of the glycolysis pathway, which are also confirmed by the Keio collection. One of them, *fbaA* is not predicted to be critical for the cell growth since the gene knockout cell keeps almost all minimal active pathways producing ATPs even if we delete both *fbaB* and *fbaA*. Specifically, the knockout lost only four minimal active pathways producing ATPs (see Table I). Thus, two hypotheses come up. One is that the four lost minimal active pathways are the most important pathways in glycolysis. The other is that the essentiality is caused by the breakdown of other cell functions. However, the first hypothesis is considered not to be true by the following discussion.

As the results in the previous section show, our method predicted three out of four essential genes. Focusing on minimal active pathways lost by gene knockouts allows us to find an important part of glycolysis. Table II shows the minimal active pathways lost by each gene knockouts of *gapA*, *pgk* and *eno*. Column 1, shows reaction names from the database of Ecocyc [11]. Reactions in the glycolysis pathway are collected in the upper part of the table. Column 2, shows corresponding genes to the reaction in Column 1 that are shown in Figure 4. Columns 3 to 16 show the 14 minimal active pathways such as $p_1, \ldots, p_{14}$ disabled by the gene knockouts of *gapA*, *pgk* and *eno*. For each column, "x" denotes a reaction contained in each minimal active pathway. In the case that two reactions have the same corresponding gene, e.g., *gnd*, we show the lacking effect of the two reactions respectively for reference. All 14 pathways are producing ATPs. For instance, a minimal active pathway p_8 consists of 8 reactions corresponding to the following genes: {*fbaB, fbaA*}, *pgk*, *eno*, *pgi*, {*pfkB, pfkA*}, {*pykF, pykA*}, *gapA*, {*gpmA, gpml, ytjC*}. This pathway p_8 is known as a typical glycolysis pathway and p_5 is known as a bypass pathway using pentose phosphate pathway that is used when p_8 is not available [8]. The glycolysis pathway p_8

Table II
14 PATHWAYS, WHICH PRODUCE ATP, DISABLED BY *gapA*, *pgk* AND *eno*
Symbols "*" denote genes predicted as critical for glycolysis.

Reaction Name	Gene	14 out of 75 pathways disabled by the single gene knockout of *gapA*, *pgk* and *eno*													
		p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}	p_{12}	p_{13}	p_{14}
PGLUCISOM-RXN_a	*pgi*	x	x	x	x			x	x	x	x	x	x	x	x
6PFRUCTPHOS-RXN	*pfkB,pfkA*								x	x				x	x
F16ALDOLASE-RXN_a	*fbaB,fbaA*								x	x				x	x
GAPOXNPHOSPHN-RXN_a	*gapA**	x	x	x	x	x	x	x	x	x	x	x	x	x	x
PHOSGLYPHOS-RXN_b	*pgk**	x	x	x	x	x	x	x	x	x	x	x	x	x	x
3PGAREARR-RXN_a	*gpmA,gpmI,ytjC*	x	x	x	x	x	x	x	x	x	x			x	x
2PGADEHYDRAT-RXN_a	*eno**	x	x	x	x	x	x	x	x	x	x	x	x	x	x
PEPDEPHOS-RXN_b	*pykF,pykA*	x				x	x		x			x	x		
PHOSPHOGLUCMUT-RXN_b	*pgm*										x				
GLU6PDEHYDROG-RXN_a	*zwf*					x	x								
6PGLUCONOLACT-RXN	*pgl(ybhE)*					x	x								
6PGLUCONDEHYDROG-RXN_1	*gnd*						x								
6PGLUCONDEHYDROG-RXN_2	*gnd*					x									
RIBULP3EPIM-RXN_a	*rpe*					x	x								
RIB5PISOM-RXN_b	*rpiA,rpiB*					x	x								
1TRANSKETO-RXN_b	*tktB,tktA*					x	x								
NAD-KIN-RXN	*nadk (yfjB)*					x	x								
MANNPISOM-RXN_b	*manA*		x	x	x					x				x	x
PGLYCDEHYDROG-RXN_a	*serA*											x	x		
RXN0-305_a	*hyi*											x	x		
RXN0-5289_b	*glxR,garR*												x		
TSA-REDUCT-RXN_1	*glxR,garR*											x			
GLUCOSE-6-PHOSPHATE-1-EPIMERASE-RXN_b	*yeaD*										x				
MANNKIN-RXN_b			x	x	x					x				x	x
GKI-RXN												x	x		
RXN0-6562												x	x		
RXN0-6418_b											x				
TRANS-RXN-158					x									x	
GLUCOSE-1-PHOSPHAT-RXN	*agp*										x				
RXN0-313_a		x	x	x	x			x			x	x	x		
TRANS-RXN-157											x				
TRANS-RXN-158A				x											x
2.7.1.121-RXN								x							
TRANS-RXN-165			x							x					
MANNOSE-ISOMERASE-RXN_a				x	x									x	x
Total Number of Reactions		7	9	10	10	12	12	7	8	10	11	11	11	11	11
Total Number of Corresponding Genes		6	6	6	6	12	12	5	8	8	8	8	8	8	8

is activated mainly and the activity of the pentose phosphate pathway p_5 is minimized at normal condition, however, p_5 is maximized when p_8 is inactivated [8]. It will be difficult to detect of all gene-expressions or enzymatic activities in the cell on every conditions, instead of this issue, our minimal pathway analysis will be a new approach for understanding of biological robustness and systems.

From Table II, we also can easily read which gene knockout disable which minimal active pathways. Note that minimal active pathways are disabled even if only one of its components is lacked. For instance, *gapA*, *pgk* and *eno* are contained in all 14 minimal active pathways and it means that the knockout of each of them immediately disables all 14 minimal active pathways. Then, the following is observed: considering all minimal active pathways of glycolysis from a whole reaction database of *E. coli*, those pathways producing ATPs always include known essential genes. In other words, the importance of 14 minimal active pathways are confirmed. In addition, it is also confirmed that even if there are computationally feasible pathways, they cannot be bypass pathways in *E. coli*. We thus can expect that *E. coli* cannot survive without those 14 pathways. This assumption allows us to predict the effect of *multiple gene knockouts*. The knockout of *pgi* disables 12 pathways but the Keio collection shows that *E. coli* is still alive with this gene knockout. In this case, it is supposed that *E. coli* manages to survive with only 2 out of 14 pathways, i.e., p_5 and p_6. In other words, those remaining pathways are supposed to be used as bypass pathways. For instance, *pgi* encoding glucosephosphate isomerase gene of glycolysis pathway that transfer D-glucose 6-phosphate to D-fluctose 6-phosphate. However, pentose phosphate pathway is available as a bypass pathway from D-glucose 6-phosphate, resulting in the gene knockout slow-growth at starting MOPS24hr and same level of the final growth of the wild cell at MOPS48hr. Then, what will happen if we additionally knockout *rpe*, *zwf*, *gnd* or *pgl(ybhE)*? Each of those gene knockouts disable both p_5 and p_6. Thus, the following double knockouts disable all 14 pathways: (a) *zwf* and *pgi*, (b) *rpe* and *pgi*, (c) *gnd* and *pgi*, and (d) *pgl(ybhE)* and *pgi*. Then, we can predict that they are critical for *E. coli*. As well as the Keio collection, Nakahigashi *et al.* measured the growth rate of *E. coli* with some combinations of double gene knockouts including the above (a), (b) and (c) [2]. Table III shows their results and both double gene knockouts of (a) and (b) affect so strong that the growth rate of *E. coli* becomes less than 0.1. Thus, as well as the prediction of single gene knockouts, our prediction agrees with the biological results for (a) and (b). However, for the knockouts of (c), our prediction disagrees with it. For this issue, comparing more combinations of genes to be critical is needed and biological evaluations for them are necessary future work.

Table III
GROWTH RATE OF *E. coli* WITH DOUBLE GENE KNOCKOUTS [2]

Knockout genes	Growth Rate [2]	
	24 hours	48 hours
(a) *zwf*, *pgi*	0.033	0.070
(b) *rpe*, *pgi*	0.001	0.140
(c) *gnd*, *pgi*	0.199	0.322

As Table II shows, four minimal active pathways disabled by the knockout of *fbaA* are also disabled by the knockout of *pgi* that is not an essential gene. Thus, the first hypothesis discussed in the former part of this section is not true. A gene *nadk (yfjB)* is similar to *fbaA*. Our method predicts that this knockout does not affect the cell growth in terms of glycolysis. However, the Keio collection shows that this is an essential gene for *E. coli*. In the case of *nadk (yfjB)*, we found that this gene knockout affects other function of *E. coli*. In relation to this, we have additional experiments for amino acid generation in the following section.

D. Results for Amino Acids Generation

We also applied our prediction method to predict gene knockout effects of the cell growth in terms of amino acid biosynthesis. Since we want to involve more genes for our prediction, we particularly focus on essential amino acids for humans, whose synthesis needs more reactions than others. In the experiments, we separately constructed pathway instances, each of which consists of the following eight amino acids as a target metabolites: L-valine (VAL), L-leucine (LEU), L-phenylalanine (PHE), L-isoleucine (ILE), L-threonine (THR), L-lysine (LYS), L-tryptophan (TRP) and L-methionine (MET). In addition, to produce the above amino acids, we added the following metabolites to the source metabolites used in the glycolysis analysis: coenzyme-A and sulfite. For each of the eight amino acids, the computation time is on average 255 seconds and the longest computation time is 877 seconds.

In contrast to the result of glycolysis, we found there are 11 single gene knockouts that destroy all minimal active pathways without the limitation of z. That is, no pathway can synthesize each target on the entire metabolic pathway of *E. coli* with those single gene knockouts. Obviously, they are predicted to be critical to produce each amino acid. Table IV shows the cell growth of Keio collection. Column 1, gene, shows knockout genes predicted as critical by our prediction. Column 2, unsynthesized target, shows target amino acids, which cannot be synthesized with the knockout of the gene in Column 1. Columns 3 and 4 show the cell growth of *E. coli* after 24 hours and 48 hours, respectively. At first, the gene knockout of *nadk (yfjB)* is predicted as critical for the cell growth in terms of six amino acids biosynthesis. This result is also supported by the Keio collection. We thus consider the essentiality of *nadk (yfjB)* to be caused by its knockout effect in amino acids biosynthesis rather than glycolysis. Table IV also shows that our method predicts

Table IV
CRITICAL GENE KNOCKOUTS FOR AMINO ACIDS BIOSYNTHESIS

Gene	Unsynthesized Target	Keio Collection [6]	
		MOPS24hr	MOPS48hr
wild	-	0.219-0.392	0.216-0.480
folE	MET	N.A.	N.A.
nadk (yfjB)	VAL, LEU, THR, ILE, LYS, MET	N.A.	N.A.
thrC	THR	0.000	0.000
thrB	THR	0.004	0.010
glnA	TRP, MET	0.005	0.015
aroC	TRP, PHE, TRP	0.009	0.020
lysA	LYS	0.012	0.021
aroB	PHE, TRP, MET	0.010	0.032
leuA	LEU	0.026	0.034
folP	MET	0.283	0.293
metH	MET	0.357	0.509

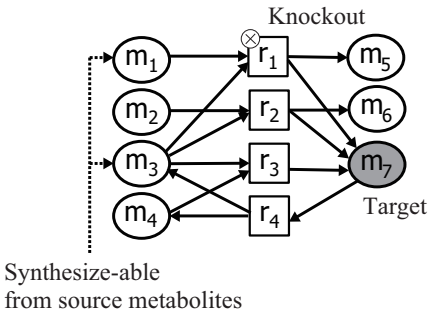
that no way to produce target metabolites with each single gene knockout: *folE*, *thrC*, *thrB*, *glnA*, *aroC*, *lysA*, *aroB*, and *leuA*. However, except *folE* and *thrC*, the Keio collection shows that *E. coli* survives with very low cell growth. One explanation for the results is that they are suspected to keep living by consuming unsynthesized amino acids from other individual cells. In this case, since the amino acids cannot be sustainably produced, those genes are recognized to be almost essential for *E. coli*.

Furthermore, the result of the Keio collection shows that the knockouts of *folP* and *metH* are not critical, although our method predicts them to be critical. We have detailed discussions on those gene knockouts in the following section.

E. Discussion for Amino Acids Generation

The difference between *folP* and *metH* in terms of amino acid biosynthesis also introduces interesting issues. At first, we consider *metH*, which constructs an enzymatic reaction methionine synthase. Its conversion is as follows: 5-methyltetrahydrofolate + L-homocysteine = tetrahydrofolate + L-methionine.

In both KEGG and EcoCyc databases, two alternative reactions exist to the above reaction. Figure 5 shows standard reaction and those two alternatives. An alternative reaction r_3, homocysteine S-methyltransferase, uses S-methyl-L-methionine (m_4) instead of 5-methyltetrahydrofolate (m_1). On the other hand, another alternative reaction r_2, 5-methyltetrahydropteroyltriglutamate–homocysteine methyltransferase, uses 5-methyltetrahydropteroyltri-L-glutamate (m_2). However, both metabolites cannot be synthesized from the source metabolites. Specifically, S-methyl-L-methionine (m_4) can be synthesized only from methionine (m_7), which is the target amino acid, and there is no reaction in the metabolic pathway of EcoCyc that can synthesize 5-methyltetrahydropteroyltri-L-glutamate (m_2), meaning that reactions are lacking in the database. The gene *folP* is on folate biosynthesis and there is no alternative in the databases. For the results of above genes, two hypotheses



r_1: 5-methyltetrahydrofolate--homocysteine methyltransferase (metH)
r_2: 5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase(metE)
r_3: homocysteine S-methyltransferase (mmuM)
r_4: homocysteine S-methyltransferase (mmuM, reverse)
m_1: 5-methyltetrahydrofolate
m_2: 5-methyltetrahydropteroyltri-L-glutamate
m_3: L-homocysteine
m_4: S-methyl-L-methionine
m_5: tetrahydrofolate
m_6: tetrahydropteroyltri-L-glutamate
m_7: L-methionine

Figure 5. Bypass Reactions for Synthesizing Methionine

are as follows: there are unknown complementary genes, or there are unknown bypasses. For the above issues, we need further analyses or researches on biological and computational level that would be creating new systems biology.

VIII. Related Work

There are several researches on metabolic pathway analyses. Schuster *et al.* proposed a method based on elementary mode analysis [15]. They focused on metabolic flux distributions corresponding to sets of reactions in metabolic pathways. A different point from our method is that their approach needs to define source metabolites strictly with a fixed amount that must be consumed in flux. In contrast, our method treats them as candidates that will be utilized; thus, we can flexibly give source metabolites. Handorf *et al.* proposed the *inverse scope problem* [16]. This is the problem of finding necessary source metabolites from target metabolites. The two differences between their problem and our proposal are as follows. One is that they only computed the cardinality minimal solution. Unlike their approach, we can generate subset minimal solution by minimal model generation. Another one is that each of their solutions includes all reactions, which are activatable from source metabolites needed to generate target metabolites. For instance, if there are two ways to produce a metabolite from source metabolites then both are mixed in one solution, that is, we cannot distinguish between them. On the other hand, our method can distinguish between the two ways, and we think that it is important to identify functionally minimal active pathways. Schaub and Thiele applied answer set programming (ASP) to solve the inverse scope problem [17], while we use propositional encoding and minimal model generation to compute minimal active pathways.

There are other researches using ASP [18], [19]. The literature [19] consider the most likely states of a reaction network with respect to given constraints and signaling pathways are analyzed with ASP. In [18], Ray *et al.* report a method using ASP to compute the steady states of a given pathway and complete lacking reactions. Unlike their approach, we use minimal model generation to compute essential reactions to produce target metabolites.

Küffer *et al.* report an approach using a Petri net [20]. Although their approach considers producibility and activatability of metabolites and reactions, they do not consider subset minimality of solutions.

IX. Conclusion and Future Work

In this paper, we propose a method to predict gene knockout effects by enumerating minimal active pathways. We formalize the extended pathway and show the definition of minimal active pathways on it. In addition, we present a computation method for the prediction. An advantage of our method is that it allows us to trace the reason for the prediction results, e.g., we can suggest the reason for the essentiality of three genes in the glycolysis pathway. This is an important feature that other methods do not have.

In the experiments, we applied our method to extended pathways of *E. coli* and made comparisons using the Keio collection. For the prediction of the knockout of 61 genes in the glycolysis pathway, our method predicted three essential genes, which correspond to the results of the Keio collection. Moreover, we analyze lethal 14 minimal active pathways and predict lethal pairs of gene knockouts, which also agree with the result of the literature [2]. In addition to the experiments in glycolysis, we found three essential genes and six almost essential genes in amino acids biosynthesis. We also discuss the reason for the difference between our prediction and results of the Keio collection with regard to the knockout of *metH* and *folP*. Although we treat relations between genes and enzymatic reactions that have one-to-one relations, we intend to extend them to relations that are more complex such as multiple relations and consider interactions among genes. Following that, we plan to apply our method to other organisms such as mice. In addition to *E. coli*, mice are well known model organisms for human study, and information available on them has been accumulated in the last decade. In particular, chromosome substitution strains are used to reveal the function of genes [21]. In addition to gene knockouts, we could adapt our method to such strains. Although there

is a large difference between *E. coli* and mice, the basic metabolism is same. This fact tells us that our method can also be a potential prediction method for mice.

ACKNOWLEDGMENTS

This research is supported in part by the 2008-2011 JSPS Grant-in-Aid for Scientific Research (A) No. 20240016, the research project of Systems Biology of Genetic Function, and the Transdisciplinary Seeds Research 2011 in the Transdisciplinary Research Integration Center. The authors would like to thank Kenji Nakahigashi and anonymous reviewers for their helpful comments.

REFERENCES

[1] T. Soh, K. Inoue, T. Baba, T. Takada, and T. Shiroishi, "Predicting gene knockout effects by minimal pathway enumeration," in *The 4th International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies (BIOTECHNO 2012)*, 2012, pp. 11–19.

[2] K. Nakahigashi, Y. Toya, N. Ishii, T. Soga, M. Hasegawa, H. Watanabe, Y. Takai, M. Honma, H. Mori, and M. Tomita, "Systematic phenome analysis of *Escherichia coli* multiple-knockout mutants reveals hidden reactions in central carbon metabolism," *Molecular Systems Biology*, vol. 5, no. 306, pp. 1–14, 2009.

[3] H. D. Jong, "Modeling and simulation of genetic regulatory systems: A literature review," *Journal of Computational Biology*, vol. 9, pp. 67–103, 2002.

[4] M. Terzer, N. D. Maynard, M. W. Covert, and J. Stelling, "Genome-scale metabolit networks," *Systems Biology and Medicine*, vol. 1, no. 3, pp. 285 – 297, 2009.

[5] C. J. Tomlin and J. D. Axelrod, "Biology by numbers: mathematical modelling in developmental biology," *Nature Reviews Genetics*, vol. 8, no. 5, pp. 331 – 340, 2007.

[6] T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori, "Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection," *Molecular Systems Biology*, vol. 2, no. 2006.0008, 2006.

[7] H. Mizoguchi, H. Mori, and T. Fujio, "*Escherichia coli* minimum genome factory," *Biotechnology and Applied Biochemistry*, vol. 46, no. 3, pp. 157–167, 2007.

[8] N. Ishii, K. Nakahigashi, T. Baba, M. Robert, T. Soga, A. Kanai, T. Hirasawa, M. Naba, K. Hirai, A. Hoque, P. Y. Ho, Y. Kakazu, K. Sugawara, S. Igarashi, S. Harada, T. Masuda, N. Sugiyama, T. Togashi, M. Hasegawa, Y. Takai, K. Yugi, K. Arakawa, N. Iwata, Y. Toya, Y. Nakayama, T. Nishioka, K. Shimizu, H. Mori, and M. Tomita, "Multiple high-throughput analyses monitor the response of *E. coli* to perturbations," *Science*, vol. 316, no. 5824, pp. 593–597, 2007.

[9] T. Soh and K. Inoue, "Identifying necessary reactions in metabolic pathways by minimal model generation," in *PAIS 2010, Proc. ECAI 2010*, 2010, pp. 277–282.

[10] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "KEGG for integration and interpretation of large-scale molecular data sets," *Nucleic Acids Research*, 2011.

[11] I. M. Keseler, J. Collado-Vides, A. Santos-Zavaleta, M. Peralta-Gil, S. Gama-Castro, L. Muiz-Rascado, C. Bonavides-Martinez, S. Paley, M. Krummenacker, T. Altman, P. Kaipa, A. Spaulding, J. Pacheco, M. Latendresse, C. Fulcher, M. Sarker, A. G. Shearer, A. Mackie, I. Paulsen, R. P. Gunsalus, and P. D. Karp, "EcoCyc: a comprehensive database of *Escherichia coli* biology," *Nucleic Acids Research*, vol. 39, no. suppl 1, pp. D583–D590, 2011.

[12] N. Eén and N. Sörensson, "An extensible SAT-solver," in *Proc. the 6th International Conference on Theory and Applications of Satisfiability Testing (SAT 2003)*, 2003, pp. 502–518.

[13] M. Koshimura, H. Nabeshima, H. Fujita, and R. Hasegawa, "Minimal model generation with respect to an atom set," in *Proc. the the 7th International Workshop on First-Order Theorem Proving (FTP '09)*, 2009, pp. 49–59.

[14] G. P. Ferguson, S. Totemeyer, M. J. MacLean, and I. R. Booth, "Methylglyoxal production in bacteria: suicide or survival?" *Archives of Microbiology*, vol. 170, no. 4, pp. 209–218, 1998.

[15] S. Schuster, D. A. Fell, and T. Dandekar, "A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks," *Nature Biotechnology*, vol. 18, pp. 326–332, 2000.

[16] T. Handorf, N. Christian, O. Ebenhöh, and D. Kahn, "An environmental perspective on metabolism," *Journal of Theoretical Biology*, vol. 252, no. 3, pp. 530 – 537, 2008.

[17] T. Schaub and S. Thiele, "Metabolic network expansion with answer set programming," in *Proc. the 25th International Conference on Logic Programming (ICLP 2009)*, 2009, pp. 312–326.

[18] O. Ray, K. E. Whelan, and R. D. King, "Logic-based steady-state analysis and revision of metabolic networks with inhibition," in *Proc. the 2nd International Conference on Complex, Intelligent and Software Intensive Systems (CISIS 2009)*, 2009, pp. 661–666.

[19] O. Ray, T. Soh, and K. Inoue, "Analyzing pathways using asp-based approaches," in *Algebraic and Numeric Biology: Proceedings of the 2010 International Conference (ANB'10)*, ser. LNCS, vol. 6479, 2011, pp. 167–183.

[20] R. Küffner, R. Zimmer, and T. Lengauer, "Pathway analysis in metabolic databases via differetial metabolic display (DMD)," in *German Conference on Bioinformatics*, 1999, pp. 141–147.

[21] T. Takada, A. Mita, A. Maeno, T. Sakai, H. Shitara, Y. Kikkawa, K. Moriwaki, H. Yonekawa, and T. Shiroishi, "Mouse inter-subspecific consomic strains for genetic dissection of quantitative complex traits," *Genome Research*, vol. 18, no. 3, pp. 500–508, 2008.

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, ENERGY, COLLA, IMMM, INTELLI, SMART, DATA ANALYTICS

issn: 1942-2679

International Journal On Advances in Internet Technology

ICDS, ICIW, CTRQ, UBICOMM, ICSNC, AFIN, INTERNET, AP2PS, EMERGING, MOBILITY, WEB

issn: 1942-2652

International Journal On Advances in Life Sciences

eTELEMED, eKNOW, eL&mL, BIODIV, BIOENVIRONMENT, BIOGREEN, BIOSYSCOM, BIOINFO, BIOTECHNO, SOTICS, GLOBAL HEALTH

issn: 1942-2660

International Journal On Advances in Networks and Services

ICN, ICNS, ICIW, ICWMC, SENSORCOMM, MESH, CENTRIC, MMEDIA, SERVICE COMPUTATION, VEHICULAR, INNOV

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, IMMM, MOBILITY, VEHICULAR, DATA ANALYTICS

issn: 1942-2628

International Journal On Advances in Systems and Measurements

ICQNM, ICONS, ICIMP, SENSORCOMM, CENICS, VALID, SIMUL, INFOCOMP

issn: 1942-261x

International Journal On Advances in Telecommunications

AICT, ICDT, ICWMC, ICSNC, CTRQ, SPACOMM, MMEDIA, COCORA, PESARO, INNOV

issn: 1942-2601