eKNOW 2012

The Fourth International Conference on Information, Process, and Knowledge Management

ISBN: 978-1-61208-181-6

January 30- February 4, 2012

Valencia, Spain

eKNOW 2012 Editors

Jaime Lloret Mauri, Polytechnic University of Valencia, Spain

Pascal Lorenz, University of Haute Alsace, France

eKNOW 2012

Forward

The fourth edition of the International Conference on Information, Process, and Knowledge Management (eKNOW 2012) was held in Valencia, Spain, on January 30th – February 4th, 2012. The event was driven by the complexity of the current systems, the diversity of the data, and the challenges for mental representation and understanding of environmental structure and behavior.

Capturing, representing, and manipulating knowledge was and still is a fascinating and extremely useful challenge from both theoretical and practical perspective. Using validated knowledge for information and process management and for decision support mechanisms raised a series of questions the eKNOW 2012 conference was aimed at.

eKNOW 2012 provided a forum where researchers were able to present recent research results and new research problems and directions related to them. The topics covered aspects from knowledge fundamentals to more specialized topics such as process analysis and modeling, management systems, semantics processing and ontology.

We take this opportunity to thank all the members of the eKNOW 2012 Technical Program Committee as well as the numerous reviewers. The creation of such a broad and high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to the eKNOW 2012. We truly believe that, thanks to all these efforts, the final conference program consists of top quality contributions.

This event could also not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the eKNOW 20102 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that eKNOW 2012 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in knowledge management research.

We also hope the attendees enjoyed the beautiful surroundings of Valencia, Spain.

eKNOW 2012 Chairs
Jaime Lloret Mauri, Polytechnic University of Valencia, Spain

eKNOW 2012

Committee

eKNOW 2012 General Chair

Jaime Lloret Mauri, Polytechnic University of Valencia, Spain

eKNOW 2012 Technical Program Committee

Gil Ad Ariely, California State University (CSU), USA / Interdisciplinary Center (IDC) – Herzliya, Israel
Werner Aigner, Institute for Application Oriented Knowledge Processing – FAW / University of Linz, Austria
Panos Alexopoulos, IMC Technologies SA - Chalandri, Hellas
Amin Anjomshoaa, Vienna University of Technology, Austria
Zbigniew Banaszak, Warsaw University of Technology, Poland
Ladjel Bellatreche, LISI- ENSMA/ Poitiers University, France
Peter Bellström, Karlstad University, Sweden
Jorge Bernardino, Polytechnic Institute of Coimbra, Portugal
Yaxin Bi, University of Ulster - Jordanstown, UK
Sabine Bruaux, Picardie Jules Verne University, France
Martine Cadot, University of Nancy1, France
Massimiliano Caramia, University of Rome "Tor Vergata", Italy
Dickson K.W. Chiu, Dickson Computer Systems, Hong Kong
Susan Gauch, University of Arkansas, USA
Olivier Gendreau, École Polytechnique de Montréal, Canada
Conceição Granja, Universidade do Porto, Portugal
Manfred Grauer, University of Siege, Germany
Pierre Hadaya, ESG UQAM, Canada
Richard Hussey, University of Reading, UK
Khaled Khelif, EADS- Val de Reuil, France
Daniel Kimmig, Karlsruhe Institute of Technology (KIT), Germany
Marite Kirikova, Riga Technical University, Latvia
Agnes Koschmider, KIT, Germany
Andrew Kusiak, The University of Iowa, USA
Franz Lehner, University of Passau, Germany
Hiep Luong, University of Arkansas, USA
Dirk Malzahn, OrgaTech GmbH, Germany
Marco Mevius, HTWG Konstanz, Germany
Roy Oberhauser, Aalen University, Germany
Daniel O'Leary, University of Southern California, USA
Andreas Papasalouros, University of the Aegean – Samos, Greece
Kenji Saito, Keio University, Japan
Erwin Schaumlechner, Tiscover GmbH - Hagenberg, Austria
Tim Schlüter, Heinrich Heine University Düsseldorf, Germany
Jan Sefranek, Comenius University, Bratislava, Slovakia

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Intranet 2.0 Based Knowledge Production

An Exploratory Case Study on Barriers for Social Software

Doris Riedl, Fritz Betz
Dept. Information Technology and Information Studies
University of Applied Sciences Burgenland
Eisenstadt, Austria
e-mail: doris.riedl@fh-burgenland.at, fritz.betz@fh-burgenland.at

Abstract— **The evolution of static intranets to dynamic web 2.0 based information systems is one way to provide space for the collaborative production of knowledge within an enterprise. Despite the fact that social software is now commonly provided for intra-company usage, this usage is below expectations in many cases. This paper, based on an exploratory case study in an international bank, shows the drawbacks as well as the drivers for the participative generation of knowledge using web 2.0 tools within an intranet. The findings, against the background of recent technology-oriented research, are three groups of possible barriers which are intertwined and therefore influence each other, namely organisational, cultural and technological barriers. Above all, the results of the case study suggest it is less meaningful to discuss if and how social software may or may not change organisations but to interpret the findings in a social science-based framework by taking the work of Boltanski & Chiapello and their understanding of the new forms of work organisation into consideration. This interpretation, while preliminary, suggests that employees using Web 2.0 software for knowledge production struggle with the ambiguity between the demands of these new forms of work and the existing, traditional organisational structures.**

Keywords - Intranet 2.0; Collaboration; Knowledge Production; Barriers; Enterprise 2.0

I. INTRODUCTION

Implementing interactive Web 2.0 based software for organisation internal usage is often accompanied by diffuse expectations, such as better knowledge management or increased productivity. However, current data shows that investment in collaboration software in many cases does not fulfill these intentions, as the usage of the tools is below expectations [1]. Nevertheless, social software based on web 2.0 principles [2] [3] is widespread in enterprises [4] [5] [6] [7] [8]. Placing the focus on the internal usage of web 2.0 based software, enterprises are now leaving the 1.0 era of intranets and turning to social intranets, providing blogs, wikis and features for social networking, such as user profiles, activity streams and microblogging [9]. Such intranet 2.0 platforms [10] are to be used by employees for information exchange, communication, networking, coordination and the collaborative production of knowledge. Driven by an IT industry hype, these projects focus on currently discussed Enterprise 2.0 concepts such as open communication, open information access, enhanced cross-departmental collaboration and open innovation. However, the realisation of these aspects is below expectations [8]. Intranet 2.0 in the above sense is a subset of Enterprise 2.0 aiming at the organisation´s internal communication and collaboration. Therefore findings of the current Enterprise 2.0 discussion are highly relevant for intranet 2.0 projects.

The term Enterprise 2.0 was proposed by McAfee [11] as "*the use of emergent social software platforms within companies, or between companies and their partners and customers*". This marked the beginning of a lively and still ongoing discussion among researchers as well as practitioners about how enterprises may benefit from the usage of social media. The current discussion concerning "*the deep impact on organisational and cultural changes*" of Enterprise 2.0 projects [12] considers the possible changes in the ways people communicate, share information, contribute and make decisions, due to the new active role of the users. But unfortunately these discussions are characterised by a lack of specific results and diffuseness. To date, there is little research into the interplay between the success rate of implementing Enterprise 2.0 initiatives, the organisation of work that manifests itself in the form of organograms, business process descriptions and standards within companies, and finally, norms and values which are rooted in a company's corporate culture.

The present paper, therefore, aims to fill part of this gap by presenting a case study on an intranet 2.0 project in an international bank: We analyse *what the potential barriers as well as the potential drivers for the collaborative production of knowledge using social software are*. Against the background of technology and business oriented research, we chose a more human-centred approach considering soft factors such as norms and values, attitudes and organisational paradigms that are all reflected in the rules and standards of the organisation.

The paper is divided into five parts. In the introduction, the context is established and the problem as a gap is addressed. Section two begins by outlining the theory underpinning the research, and discusses how the term collaboration is embedded in the Enterprise 2.0 discussion. This part then reviews the literature concerning the organisational and cultural aspects of collaboration using social software and also considers a social science

 ISBN: 978-1-61208-181-6

perspective on work organisation in general. Since a substantial part of Enterprise 2.0 empirical research is documented in cases studies [13], we chose a case study approach, too. Our case study design and our research method in detail are described in section three. Section four focuses on the specific drivers and barriers as a result of the case study followed by a discussion of the results. The fifth and last section, the conclusion, discusses the consequences for future intranet 2.0 projects and includes an outlook for further research.

II. COLLABORATIVE KNOWLEDGE PRODUCTION WITHIN THE INTRANET 2.0

The amount of related literature that has been published on collaborative knowledge production within intranet 2.0 is very limited. For this reason, the more general question about whether the use of social software is adequate for knowledge workers to generate new knowledge was used as a starting point for the literature search: Levy [3] investigated how knowledge management could and should be enhanced in light of the Web 2.0 and states that the principles of Web 2.0 and Knowledge Management are very similar. Finally, she recommends adopting the participative nature of Web 2.0 for production and sharing of knowledge in organisations and suggests starting with wikis and blogs. Studies at Fraunhofer ISST [14] show that the use of Web 2.0 in enterprises has its biggest impact on knowledge work, innovation and cooperation, although its potential is hardly exploited. Paroutis and Al Saleh [15] likewise state that blogs, wikis and other social software have distinct technical features that foster knowledge sharing. Stocker [16] as well shows in an empirical study that knowledge transfer within enterprises profits from wikis and blogs. Also current data on usage of Web 2.0 software in enterprises has shown benefits for knowledge management such as increasing speed of access to knowledge [4] and more efficient usage of explicit and tacit knowledge [6].

Looking closer at the development of "*collaboration*" as a term for people working together to reach a common goal, we find a close connection to the SLATES-concept of McAfee [11] that was extended to FLATNESSES by Hinchcliffe [17]. SLATES is an acronym for Search, Links, Authorship, Tags, Extensions, Signals and was created to provide a basic concept for Enterprise 2.0 software. The extensions added by Hinchcliffe [17] include Freeform, Network-oriented, Social, and Emergence. Software providing these features and characteristics empowers users to participate in and contribute actively to the information flow inside an enterprise and also across organisational borders, in the same way as they are used by social software platforms in the public web. By these means, Web 2.0 principles such as transparency, accessibility and personalisation can find a way into an organisation. To sum up, electronic collaboration (E-Collaboration) can be seen as a special case of IT-supported cooperation to achieve a common goal with shared responsibility for the results [18] using software with its nucleus in Enterprise 2.0 strategies: *Authoring* (access to platforms to produce one's own content), *Social* (not hierarchical, transparent) and *Network-oriented* (Web-based, addressable and reusable content).

Enterprise 2.0 software enables freeform collaborative production of content without an imposed structure such as predefined business processes or hierarchical access rights [19]. According to Schachner and Tochtermann [20], the internal use of Web 2.0 software requires and/or leads to changes in the way people work together: self organisation instead of top down coordination in terms of spontaneous, mostly voluntary cooperation; open information flow instead of secretly working task forces; trust and openness to criticism instead of sanctioning mechanisms; and individual responsibility for "pulling" the necessary information. This goes hand in hand with a change in the mindset of the users, namely thinking in business models and solutions instead of concentrating on the technology.

How Social Software Will Change the Future of Work is not only the subtitle of Cook´s book on Enterprise 2.0 [21] but also one of the central questions discussed in the Enterprise 2.0 community. On the one hand, the discourse is dominated by business consultants and analysts such as Don Tapscott [22], Dion Hinchcliffe [9] [17], Andrew McAfee [11] [19] and Niall Cook [21], many of who argue that social software is a driver of organisational and cultural changes. They believe that giving employees the technical possibility to collaborate eventually initiates the transformation of enterprises into, to some degree, non-hierarchical, self-organised networked organisations with an open culture. On the other hand, authors with a knowledge management perspective suggest that an appropriate organisation and culture is a prerequisite for E-Collaboration rather than a consequence: Davenport [23] states in his blog that "*the absence of participative technologies in the past is not the only reason that organisations and expertise are hierarchical*". Schneckenberg [24] also argues that organisational factors, such as adequate decision-making policies, corporate governance and value systems ingrained in the corporate culture, are preconditions for the acceptance and sustainable use of Web 2.0 technologies in companies. These findings suggest that flat hierarchies and transparency, either as a prerequisite or a consequence, are closely connected with successful E-Collaboration using social software as internal tools.

For this reason, it seems useful to look closer at the aims and objectives behind fostering collaborative work and to also take the impact of the work organisation into consideration. In their major work "The New Spirit of Capitalism" [25], Boltanski and Chiapello reviewed management literature that influenced the thinking of executives and employees of companies over the last decades. They argue that the hierarchical Fordist work structure was abandoned from the middle of the 1970s onwards and a new network-based form of organisation came into existence. This new form of work organisation is founded on employee initiative and work autonomy.

According to Boltanski and Chiapello [25], the "new spirit of capitalism" means self-fulfillment as a strategy to mobilise labour. The new highly flexible work force does not separate social life into a private and a professional part, but

lives and acts in a networked world with multitudinous contacts in projects as the main organisational unit. Everything can be a project – the construction or the closedown of a plant, the reorganisation of a company or a play in a theater [25].

The new work force is characterised by intrinsic motivation, self-organised effort, autonomy, self-management, spontaneity, and communicative competence using social media. People are either self-employed (micro-enterprises) or work as an employee competing in internal markets, with teamwork being highly important due to the rising numbers of projects and project like-tasks. These new work structures demand high flexibility and mobility together with permanent reachability.

In this new form of work organisation *activity* is the new norm to measure the value of people and objects. *Activity* means starting projects and contributing actively to projects while using networks for contacting and getting information to eventually initiate new projects [25]. Consequently, the traditional norms *efficiency* and *properly executed actions* have been replaced.

III. CASE STUDY

In knowledge management research, the case study method is often applied since it has broad applicability. Hence, there are different kinds of case studies depending on the underlying research design [26] [27].

A. Case Study Design

As mentioned before, little research has been done into how work organisation, a company's corporate culture and the success rate of implementing Enterprise 2.0 initiatives interact with each other. In this early stage of research, when "*…"how" questions are posed, the investigator has little control over events and the focus is on a contemporary phenomenon within a real-life context*" [26], case studies are the preferred method. Yin [26] differentiates between three basic types: exploratory, explanatory, and descriptive case studies. Each of these approaches can either be single or multiple case studies.

Our case was a single case study examined using an exploratory design, as data was first collected and then patterns in the data were identified. To achieve a more abstract view, the identified patterns were put in a theory based frame and a more general model was derived.

B. The Case

In May 2009, an international financial services provider based in Austria decided to build up a new intranet. The aim was to provide up-to-date information for the employees as well as to enhance collaborative work. In the following year, a pilot project using Microsoft® SharePoint® Server 2007 as a platform was implemented and more social software applications for rating, commenting, communicating in forums and wikis were added. One major task of the intranet 2.0 project was the implementation of so-called "topic areas" on the SharePoint® Server. These areas were intended to be managed and used by specific employees, called "topic coordinators". These were highly skilled specialists, e.g., software development specialists, who were responsible for the production and the enterprise-wide distribution of topic-specific knowledge. The new topic areas were designed as future places for participative production and allocation of topic knowledge under the lead of the coordinators. The whole intranet 2.0 initiative was seen as a first step to becoming an Enterprise 2.0.

In the pilot phase, about 200 intranet users were invited to participate. Although there were plenty of internal marketing activities for the new intranet, the usage of the new platform, measured by key performance indicators produced by the SharePoint® Server, e.g., usage statistics, fell short of expectations. In particular, the project leader was not satisfied with the low activity of users participating in the topic areas. The idea of providing a well-designed platform to enable employees to participate and give their input and comments on various topics simply did not work as intended.

At this point of time, the authors of this paper were invited to analyse the situation, identify what was causing the unsatisfying key performance indicators and propose improvement measures based on the findings.

C. Research Method

First, the project leader of the bank was interviewed using a problem-centred interview technique to give orientation and facilitate generation of first assumptions. Based on the received data, a focus group consisting of selected topic coordinators was established, who also represented the users of the platform. A set of workshops with this group was designed to undertake an in-depth analysis of the situation. Finally, three workshop sessions moderated by the authors of this paper took place. Each workshop was followed by discussions within the research team and first assumptions about the roots of the identified barriers were generated. These assumptions were subject to a deeper, theory-based analysis and were again reflected on with the members of the focus group in the next workshop session.

IV. FINDINGS

The researchers placed their main focus on identifying the barriers or drawbacks for the unsatisfying usage of the topic areas. However, the drivers for the collaborative knowledge production were also discussed: The members of the focus group identified the topic areas used as a central knowledge repository with a well structured file sharing and good search function as helpful. This knowledge repository was a storage for longer existing documents that had already passed a quality assurance process. Therefore, all employees who had access to this repository could be sure to get information that was up-to-date and confirmed by management. As a consequence, employees had less need to call the topic coordinators by phone in case of a specific question or to send them an email that induced a time relief for the specialists. Another factor influencing the usage of the new topic areas positively was the possibility to get automatic alerts when a document was changed. Nevertheless, the documents in the repositories were mostly static files with no need to update frequently.

 ISBN: 978-1-61208-181-6

A significant part of the analysis of the drawbacks was discussing the underlying causes. Consequently, three types of barriers were identified, each characterized by being embedded in the same context:

- *barriers rooted in the organisational culture* (in the values and behavioural norms of the organisation)
- *barriers rooted in the organisation itself* (in the organisational structure and the business processes)
- *barriers rooted in the technology* (in the implementation of applications).

The following paragraphs explain the three barrier types as identified in the case study in detail followed by a discussion of the findings.

A. *Barriers rooted in the organisational culture*

Each organisation has its own internal values and often unspoken behavioral norms that may be contrary to Enterprise 2.0 paradigms such as open communication, self organisation or decentralized decisions. In the explored case study, the norm "valid knowledge has to pass a certain quality assurance process" was dominant due to compliance requirements of a financial service provider. Not surprisingly, this was one of the underlying reasons for the unsatisfying usage of the topic areas in the intranet 2.0. Furthermore, the identified attitudes regarding a "no blame organisation" were different among employees of the bank: some believed in the participation of many to produce knowledge of high quality (wisdom of crowds concept) whereas some adhered to the traditional belief in the expertise of a few, highly-skilled specialists. As part of the organisational culture, knowledge was seen as something owned by the organization that should be distributed within the organisation only carefully (similar to a company secret). Even IT knowledge that was found on the public web was affected by this approach.

Furthermore, E-Collaboration and participation in knowledge production need an organisational culture, where self-organisation and sharing is desirable. The traditional hierarchical culture of the financial service provider in our case study, based on divisions and command and order, allowed only little room for acting autonomously and collaborating beyond the daily routine. In particular, the understanding and expectations of E-Collaboration varied between management, project management and the users.

B. *Barriers rooted in the organisation itself*

Among other things, an organisation is manifested in the structure or hierarchy of an enterprise. In the analysed case, the hierarchy of the organisation, represented in divisions, departments and sub-departments, was mapped in the intranet 2.0 applications and the corresponding access rights, thus restricting E-Collaboration and participatory production of content. But, what was most important, there was no organisational link between the daily work in the business processes and the content generation in the topic areas. In addition, employees lacked time to work on the topic areas besides their daily routine. The job descriptions of employees were not updated to accommodate the new tasks and responsibilities resulting from the usage of the new intranet. The internal organisational rules left marginal space for knowledge production besides the highly standardised quality assurance processes; there was nearly no possibility for open, dynamic and up-to-date ad-hoc generation and usage of content. It was also still unclear and under discussion who the potential users were and what usage the content stored in the topic areas was intended for.

The former role of the topic coordinators before starting the intranet 2.0 project was that of a specialist collecting information, generating and distributing knowledge within the organisation (one-to-many communication). In the discussion with the coordinators, we found different attitudes about if and how this role had to be changed to foster participation of the other employees in knowledge production (many-to-many communication). One of the topic coordinators made the following point: he called the new role "E-Collaboration animator" expressing the feeling that in future his expertise may not been seen as his valuable skill but will be replaced by the need to be an experienced moderator of online communities. This statement suggests that the new role of the "proactive knowledge provider" was unclear and not defined in an organisational context.

C. *Barriers rooted in the technology*

The technology itself, in our case the SharePoint® Server plus the implemented add-on applications such as a wiki, a search function and a forum software, worked quite well. The identified technical barriers were some constraints such as a cumbersome document upload function with many compulsory tags for classification, and the search function lacking a document preview. Low usability due to extensive menu structures and slow performance were also commented on by the members of the focus group. As most content was stored in documents, users could not make quick ad-hoc updates of the information, i.e. the documents had to be down and uploaded to be changed. The members of the focus group also stated that in some knowledge areas on the SharePoint® Server there were either too many documents or too few. To summarise, the intranet 2.0 was not seen as the primary source of knowledge; the users still preferred to search on the internet.

D. *Discussion*

It is interesting to note that in our case barriers rooted in the organisation itself and in the culture seemed to have a greater impact on the collaborative production of knowledge in the topic areas than the ones caused by low usability or low functionality of the software. Consequently, initiatives to facilitate adoption of social software, such as training, project marketing, working with key users, getting management to use the tools actively by themselves, are not sufficient. In interpreting these findings, we have to consider the previous research of Pan and Scarbrough [28] that was undertaken already two decades ago. They developed a theoretical model with three major layers that were required for technological innovations (in their case, a knowledge management system) to be successful: *Infrastructure* (the hardware/software that enables the physical/ communicational contact between network members),

 ISBN: 978-1-61208-181-6

Infostructure (the formal rules, that govern the exchange between the participants in the network) and *Infoculture* (the stock of background knowledge that actors take for granted and that is embedded in the social relations surrounding work group processes). Pan and Scarborough called the latter also the cultural knowledge that defines constraints on knowledge and information sharing. Most importantly, their conclusion is that knowledge management systems *"…involve more than technology but rather a culture in which new roles and constructs are created. It changes the communication patterns between individuals and teams, and also alters the design of the organisation by fostering new processes and structures*" [28]. Interestingly, the three types of possible barriers we found in our case study seem to match in some way the three layers of Pan and Scarborough´s model. However, in our case (a social software based intranet) the effects of a technological innovation as stated by Pan and Scarborough could not be observed and there were no signs for alterations in the organisation. Consequently, it appears that the identified barriers rooted in the (knowledge) culture turned out to be the major constraints.

In addition, our findings support the view that all three types of barriers are intertwined and therefore influence each other. In particular, the organisation itself - in a sense the actual business processes and the organisational structure - is mirrored in the IT applications and the corresponding access rights. On the other hand, the organisation itself is effected by the norms, values and paradigms of an enterprise, thus reflecting the organisational culture. For example, in our case the most important single barrier was the lack of alignment of the intranet 2.0 applications to the business process requirements. The business processes were implemented in the form of internal rules and standards. Furthermore, the organisational culture of the financial service provider in our case study may be characterised as being traditional and hierarchical and dominated by compliance requirements. Therefore, the internal standards defined an accurate quality assurance process for documents, with several confirmation steps on management level built-in. Despite this, the topic coordinators were asked to generate knowledge collaboratively which indicated the requirement to publish not-confirmed content as well. This is a clear example of how the organisational culture determined by a traditional hierarchical work organisation (division of work, control and command) may influence adoption of E-Collaboration software via business process regulations.

These findings are also in accordance with our previous discussion about the "new spirit" in work organisations according to Boltanski and Chiapello [25]. As mentioned before, the new form of work organisation is replacing *efficiency,* the traditional measurement for employees and processes, with *activity*. From our perspective, *activity* in connection to knowledge production may be seen as an autonomous behavior of users, who act on their own initiative participating voluntarily and contributing interactively, as known from the production of user provided content on the public web. Hence, we presume that employees suffer from the ambiguity between the measurement of the quality of work in the "new spirit" [25] and the measurements in traditional structures. For instance, in our case study we observed the conflict situation of the topic coordinators: On the one hand, they were part of a hierarchical organisation, had to work efficiently (which was also implemented in the Management-by-Objectives), and were obliged to stick to the internal standards and rules. On the other hand, they were requested by the intranet 2.0 project to collaborate autonomously, initiatively and spontaneously. Above all, we tend to believe that the ambiguity between the new norm *activity* and the traditional value *efficiency* is the unspoken but nevertheless underlying cause of the unsatisfying usage of the intranet 2.0 platform in our case.

V. CONCLUSION

Research into participative knowledge generation utilizing social software, especially with a focus on the organisation and its people is in the early stages. The case study presented in this paper has pinpointed some specific drivers and drawbacks for intranet 2.0 based E-Collaboration. The latter have been grouped into three types of possible barriers, each type characterized by being embedded in the same context. More empirical data will need to be gathered to justify these types.

For practitioners, the quintessence of the present paper is that activities aiming to support intranet 2.0 initiatives, especially those with focus on the collaborative production of knowledge, have to consider all three types of possible barriers. Only optimizing the underlying (information) technology is not enough - the probability of failure of the whole initiative will still prevail. For instance, the organisation itself must be prepared to empower employees to use social software appropriately, e.g., by allocating sufficient time resources or adjusting the job descriptions. Daily routine activities, i.e. those to do with the business processes must be linked to the intranet 2.0 software to ensure the intended usage. All this must be considered against the background of the internal norms and values. An organisation holding on to a strictly hierarchal culture may be a limiting factor for any intranet 2.0 initiative.

On the whole we believe it is less meaningful to discuss if and how social software may or may not change organisations but to interpret the findings in a broader social science-based framework. Taking the work of Boltanski and Chiapello and their understanding of the new forms of work organisation into consideration, namely projects and networks, we believe that employees using social software for knowledge production struggle with the ambiguity between the demands of these new forms of work and the existing, traditional organisational structures. More research is needed to better understand the consequences of this development, especially how the discrepancy between traditional "tayloristic" organisations and the expected participation of the employees in the "Enterprise 2.0" may be bridged.

 ISBN: 978-1-61208-181-6

ACKNOWLEDGMENTS

This work was funded by FFG / Österreichische Forschungsförderungsgesellschaft mbH in grant No. 821111 "eCollaboration 2.0: Collaboration Tools und Social Media für Teamarbeit in KMUs". We are grateful to Michael Zeiller for challenging discussions and helpful comments and to Veronica Dal Bianco for proofreading this paper.

REFERENCES

[1] A. Stiehler, "IT Reality Check. Collaboration zwischen Anspruch und Wirklichkeit", PA/Berlecon Report by order of Beck et al. Services, May 2011, retrieved from http://www.bea-services.de/index.php?/ger/Knowledge-Center/IT-Reality-Check on June 17, 2011.

[2] T. O´Reilly, "What Is Web 2.0? Design Patterns and Business Models for the Next Generation of Software", Tim O´Reilly´s Blog, September 30, 2005, retrieved from http://oreilly.com/pub/a/web2/archive/what-is-web-20.html?page=1 on August 5, 2011.

[3] M. Levy, "WEB 2.0 implications on knowledge management", Journal of Knowledge Management, Vol. 13 NO. 1, 2009, pp. 120-134.

[4] J. Bughin and M. Chui, "The rise of the networked enterprise: Web 2.0 finds its payday", McKinsey Quarterly, December 2010, retrieved from http://www.mckinseyquarterly.com/The_rise_of_the_networked_enterprise_Web_20_finds_its_payday_2716, on August 24, 2011.

[5] ZEW, "Interaktiv, mobil, international – Unternehmen im Zeitalter von Web 2.0", Zentrum für Europäische Wirtschaftsforschung GmbH, IKT-Report September 2010, retrieved from http://ftp.zew.de/pub/zew-docs/div/IKTRep/IKT_Report_2010.pdf, on March 2, 2011.

[6] J. Leibhammer and M. Weber, "Enterprise 2.0. Analyse zu Stand und Perspektiven in der deutschen Wirtschaft", BITKOM-Studie 2008, retrieved from http://www.bitkom.org/e/publikationen/38338_60082.aspx, on March 7, 2011.

[7] N. Scheidegger, D. Vimalassrey, Ph. Surber, and J. Richard, "Enterprise 2.0 – die kollektive Intelligenz als Wettbewerbsvorteil", study of sieber&partners 2009, retrieved from http://shop.pascal-sieber.ch/shop/de/produktdetail.asp?ProduktID=48202 on March 7, 2011.

[8] T. Petry, F. Schreckenbach, and T. Nienaber, "Wenn wir wüssten, was wir wissen", Personalwirtschaft 12/2010, pp. 50-52.

[9] D. Hinchcliffe, "Social Intranets: Enterprises grapple with internal change", Dion Hinchcliffe´s Blog - Enterprise Web 2.0, ZDNet, 2010, retrieved from http://www.zdnet.com/blog/hinchcliffe/social-intranets-enterprises-grapple-with-internal-change/1410, on April 18, 2011.

[10] H-G. Herrmann, "Internet und Intranet 2.0. Chancen und Risiken für Großkonzerne", 2007, retrieved from http://www.oscar.de/archiv/2007_02/magazin/Evolution_Web_02_01_Internet_Intranet_2.0_DAX30-Studie.pdf on February 24, 2011.

[11] A. McAfee, "Enterprise 2.0: The Dawn of Emergent Collaboration", MIT Sloan Management Review 47(3), 2006, pp. 21–28.

[12] A. Auinger, A. Hochmeier, and D. Nedbal, "Organization-driven Approach for Enterprise 2.0 Projects", Tagungsband 5. Forschungsforum der österreichischen Fachhochschulen, 2011, pp.136-139.

[13] M. Zeiller and B. Schauer, "Adoption, Motivation and Success Factors of Social Media for Team Collaboration in SMEs", Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies, 2011, not yet published.

[14] F. Fuchs-Kittowski, N. Klassen, D. Faust, and J. Einhaus, "A Comparative Study on the Use of Web 2.0 in Enterprises", Proceedings of I-KNOW ´09 and I-SEMANTICS ´09, September 2009.

[15] S. Paroutis and A. Al Saleh, "Determinants of knowledge sharing using Web 2.0 technologies", Journal of Knowledge Management, Vol. 13 NO. 4, 2009, pp. 52-63.

[16] A. Stocker, "Wikis and Weblogs im Wissensmanagement: Nutzenklassen und Erfolgsfaktoren", Proceedings of 6th Conference on Professional Knowledge Management, February 2011, retrieved from http://iwi.uibk.ac.at/download/downloads/Publikationen/wm2011_tot.pdf, on November 17, 2011.

[17] D. Hinchcliffe, "The state of Enterprise 2.0", Dion Hinchcliffe´s Blog - Enterprise Web 2.0, ZDNet, 2007, retrieved from http://www.zdnet.com/blog/hinchcliffe/the-state-of-enterprise-20/143 on March 24, 2011.

[18] K. Riemer, "eCollaboration: Systeme, Anwendungen und Entwicklungen", eCollaboration. Praxis der Wirtschafts-informatik, vol. 267, K. Riemer, and S. Strahringer, Eds., 2009, pp. 7-17.

[19] A. McAfee, "Enterprise 2.0: new collaborative tools for your organization's toughest challenges", Boston: Harvard Business Press, 2009.

[20] W. Schachner and K. Tochtermann, "Corporate Web 2.0. Band II. Web 2.0 und Unternehmen - Das passt zusammen!" Aachen: Shaker, 2008.

[21] N. Cook, "Enterprise 2.0 - How Social Software Will Change the Future of Work", Aldershot: Gower, 2008.

[22] D. Tapscott and A. Williams, "Wikinomics - how mass collaboration changes everything", New York: Portfolio, 2006.

[23] T. Davenport, "Why Enterprise 2.0 Won't Transform Organizations" , Harvard Business Review Blog, March 21, 2007, retrieved from http://blogs.hbr.org/davenport /2007/03/why_enterprise_20_wont_transfo.html on March 27, 2011.

[24] D. Schneckenberg, "Web 2.0 and the shift in corporate governance from control to democracy", Knowledge Management Research & Practise 7, 2009, pp. 234-248.

[25] L. Boltanski and E. Chiapello, "The New Spirit of Capitalism", London: Verso (french: 1999), 2007.

[26] R.K. Yin, "Case Study Research. Design and Methods", fourth Edition, Applied Social Research Methods Series Volume 5, Thousand Oaks: Sage, 2009.

[27] Eisenhardt, K., "Building Theories from Case Study Research." The Academy of Management Review 14 4, Oct. 1989, pp. 532-550.

[28] S.L. Pan and H. Scarbrough, "Knowledge Management in Practise: An Exploratory Case Study", Technology Analysis & Strategic Management Sep 1999 11 3, 1999, pp. 359-374.

 ISBN: 978-1-61208-181-6

Information and Communication Technology Infrastructure in E-maintenance

Muhammad S. Al-Qahtani
Saudi Aramco
Dhahran, Saudi Arabia
E-mail: qahtms1b@aramco.com

***Abstract – The major objective of this p*aper is to provide further insights into the Information and Communication Technology (ICT) infrastructure for supporting e-maintenance processes in today's manufacturing environments. To achieve this objective existing e-maintenance models were investigated and an appropriate model was selected based on (i) its currency, (ii) its relevance to the manufacturing industry, and (ii) an explicit role being played by the ICT for enabling various e-maintenance activities. The ICT component of the selected framework is then further expanded by identifying specific ICT component technologies that are currently available for supporting various e-maintenance activities within the framework. Therefore the major contribution of the current study includes (i) identification of an existing e-maintenance framework with explicit focus on ICT, and (ii) to purposeful review of the current ICT literature in order to identify current ICT technology components that can be used in order to support various e-maintenance activities of the selected e-maintenance model.**

Keywords – manufacturing industry; maintenance; e-maintenance; conceptual model; Information and Communication Technology (ICT)

I. INTRODUCTION

As a multidisciplinary field, e-maintenance is related to a variety of research fields ranging from operation & maintenance engineering, to software engineering, information systems, and business management [1]. As a result of such inter-disciplinary nature of e-maintenance, a variety of theoretical and research perspectives can be adopted I order to investigate the phenomena. The perspective adopted in the current study is the Information and Communication Technology (ICT) perspective.

E-maintenance addresses emerging requirements of today manufacturing industry and provides various benefits in form of increased availability, reduced lifecycle and set up cost, facilitated the integration of maintenance support technologies with existing material and personal resources, increased customer-value, continuous improvement of maintenance management, improved decision making process [1], [2], [3], [4], [5].

The current study is presented in the form of a review paper, and as a result, its findings are synthesized from the existing literature on e-maintenance and ICT domains. One major motivation for conducting the current study has been a lack of sufficient insights into the existing ICT component technologies that can be used for both supporting as well as enabling various e-maintenance activities in today's manufacturing environments. While practitioners have been busy with utilizing a variety of ICT technology components for supporting their e-maintenance activities, little academic research seem to have been conducted to provide a taxonomy-like knowledge representation of the current ICT technologies that can be used in today's manufacturing sector.

The current study extends existing work on e-maintenance by providing an ICT classification scheme for e-maintenance activities by identifying specific ICT components from the current literature that provide required support. Findings of the study which is represented in an ICT classification scheme, can in turn serve as a supplement to the selected e-maintenance model, and collectively referred to as *integrated e-maintenance architecture* incorporating ICT components and e-maintenance activities. The ICT classification scheme is presented textually in the section titled "ICT Infrastructure for E-maintenance".

II. E-MAINTENANCE STRATEGIES/PROCESSES

As a multi-disciplinary e-disciplinary research field, e-maintenance is a combination of two e-domains: 'e-manufacturing' and 'e-business' [6]. It is defined as "maintenance managed and performed by virtue of computing" [1], "it integrates ICT within the maintenance strategy to face the new challenges of supporting e-manufacturing" [7], and "provision of maintenance support services remotely with the aid of ICT" [5]. In this section a review of literature is provided to further clarify the e-maintenance side of the theoretical foundation of this study which primarily focuses on current e-maintenance strategies. The outcome of this review is identification of a recent ICT-oriented e-maintenance framework where ICT is given critical role of enabling and supporting various e-maintenance activities. Such bias in reviewing the current literature has been deliberate and consistent with the overall aim of the study in one hand, and the enhanced role of ICT in e-maintenance activities, from having a 'supporting' role to having an 'enabling' role. Below is a summary of the

 ISBN: 978-1-61208-181-6

existing e-maintenance strategies adopted by today's primarily knowledge-based manufacturing organizations:

Remote maintenance: It is based on the notion of distance and transfers data from one site to another one remotely without the physical access to the item [8].

Predictive maintenance: It is concerned mainly with detecting hidden and potential failures and predicting the condition of the equipment [9].

Real-time maintenance: Maintenance operators can respond to any situation by the real-time remote monitoring of equipment status coupled with programmable alerts [10].

Cooperative maintenance: The work is divided to independent tasks, every actor assigned to a part of the resolution of the problem and the coordination is done during the assembly of partial results [10].

Collaborative maintenance: The work is synchronized and coordinated so as to build and to maintain a common vision of the problem [11].

Preventive maintenance: The objective of preventive maintenance is to decrease the probability of failure in the time period after maintenance has been applied [12].

Corrective maintenance: Corrective maintenance strives to reduce the severity of equipment failures once they occur [13].

On the other hand e-maintenance processes have been identified and classified by Kajko-Mattsson et al. [1] and Muller et al. [11] as (i) 'diagnostics', (ii) 'prognostics', (iii) 'planning and production control', (iv) 'documentation' such as technical publications, (v) 'electronic log books and technical records', (vi) 'repair order/work order', and (vii) 'quality assurance and reliability analysis'.

All above e-maintenance strategies and processes would require support from ICT in a variety of ways. For example, the 'remote maintenance' strategy would require ICT component technologies that maintain a ubiquitous environment for the maintenance workers whereas the 'predictive maintenance' strategy would need strong ICT support in the areas of *business intelligence* and *decision support* systems and technologies. The current study provides a generic guide to the ICT component technologies without adhering to a specific strategy. This will facilitate identification of matching each of the proposed ICT component technologies with a particular strategy.

III. E-MAINTENANCE FRAMEWORK

A comprehensive architectural framework for e-maintenance has been proposed by Han and Yang [6] and is widely used by researchers in the fields of management and ICT mainly because it assigns an explicit role for ICT as an enabling factor for supporting various maintenance activities. The current study adopts this framework and elaborates on the ICT component by exploring existing ICT component technologies that can be used in conjunction with the above conceptual framework.

The framework mimics the traditional holistic maintenance shops in multi-division environments with a centralized maintenance centre and several local maintenance centers, and closely resembles the model that has been adopted by the Saudi Aramco where the author is employed.

The maintenance center is a sharable platform that interconnects research groups, experts, repair shops, and manufacturing divisions via internet and communication techniques. The local maintenance centers provide routine services to their respective manufacturing sites that do not involve ad-hoc decision-making and/or fundamental changes to the existing local facilities. In the event when such needs arise the latter uses the shared facilities of the maintenance center both for problem-solving as well as for implementing change management and supporting relevant decisions. Maintaining an effective communication and coordination activities between the local and central maintenance centers is one major role of the ICT infrastructure that have been discussed in the next section.

The lower part of Figure 1 represents the ICT infrastructure component of the architecture that supports activities within the e-maintenance framework, and is the focus of the current study. In the following section the latter part of the architecture is described in more detail through a review of the current literature. The identified ICT components have been selected from the literature on the basis of their relevance and appropriateness in relation to supporting various e-maintenance activities.

 ISBN: 978-1-61208-181-6

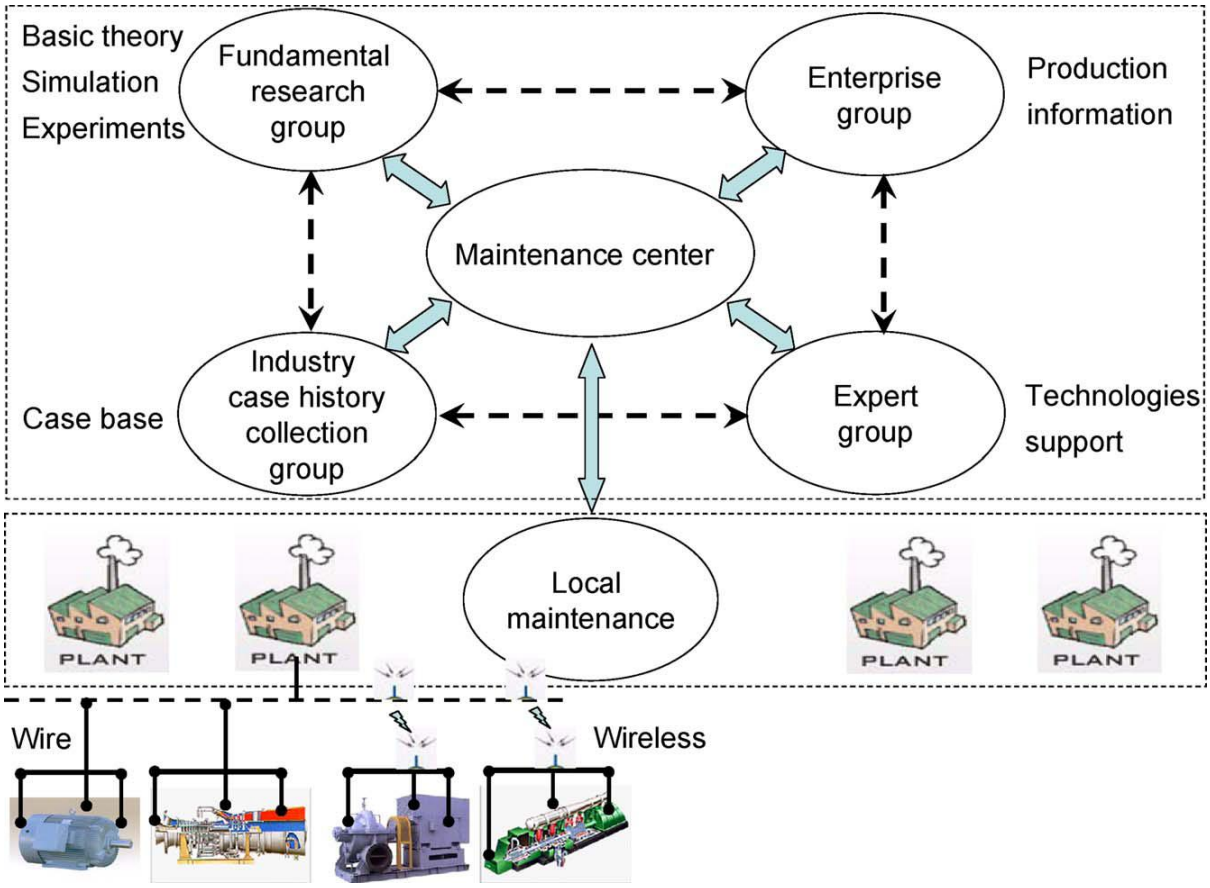


Figure 1 – Architecture of a hypothetical e-maintenance system; adopted from Han and Yang, 2006.

IV. ICT INFRASTRUCTURE

It is claimed that ICT infrastructure in e-maintenance must ensure that the level of service quality expected for the process execution is maintained for scalability and availability [14]. From the e-maintenance point of view the ICT infrastructure is composed of one or several networks with servers, workstations, applications, databases, smart sensors, PDA, and many more [14]. Furthermore, such role for ICT has also been characterized by its operating principles such as wireless infrastructure as well as deploying the right ICT related standards for presentation, storage, exchange, and process communication [15]. In the following section a summary of the most relevant categories of ICT technology components have been identified from the literature. Such categorization represents one major contribution of the current study.

- New sensors such as smart sensors—MEMS (micro-sensor technology equipped with autonomous power, memory cells, analogue amplification, converter, etc. well adapted for vibration analysis, oil analysis), wireless sensors, and sensor networks. The sensors are the main factor for performing the basic e-maintenance activities which materialize the ‘‘Condition-Based Maintenance’’ concept (CBM). Therefore, these sensors support more than conventional capacities (such as CM, diagnosis, prognosis) [15], [14].
- RFID tag (passive and active; Radio Frequency Identification Device) is used for operator and component identification, storage of conventional data, and traceability of the past maintenance actions. In addition, for it can be used for geo-localization of the maintenance tools.
- Global Positioning System (GPS) in a complementary technology to the RFID tags and is used for distinguishing location of an operator or the maintenance tools.
- Wireless technologies lead to considerable savings in networking costs and provide high degree of flexibility that are not normally provided by wired systems. Wireless Personal Area Networks (WPAN) such as IEEE 802.11, 802.15.4 ZigBee, 802.15.1 Bluetooth; Wireless Local Area Network (WLAN) such as WiFi, WiMax; GSM-UMTS (for long distance) are currently the main wireless technologies [15], [4], [17], [18].

 ISBN: 978-1-61208-181-6

- Innovative communication equipment such as "virtual reality" for supporting man/machine or man/man exchanges for speaking, hearing, seeing, touch, and feel.
- Tools for diagnostics and prognostics that support maintenance decision-making include E-CBM and remote sensing devices. These technologies are deployed for monitoring the condition and performance of physical assets [19], [20]. Furthermore, in an e-CBM-enabled environment data are transmitted through the Web to a secure site for analysis and decision making [20].
- PDA, SmartPhones, Graphic tablets, harden laptops, etc. (equipped with WiFi, Bluetooth, RFID Reader, Windows Mobile).
- Specific standards for ensuring the integration between all the IT components and e-Maintenance solutions.
- Web services (for monitoring, diagnosis, prognosis, scheduling) protocols and technical standards (Internet-based technologies) used for exchanging data between applications within heterogeneous environments: SOAP (Simple Object Access Protocol) for message exchanging; WSDL (Web Service Description Language); UDDI for referencing the web services, etc.
- Full Web-CMMS (e-CMMS) is a CMMS (Computerized Maintenance Management System) able to monitor and manage the preventive maintenance activities of the organization but by offering new functionalities such as ASP (Application Service Provider/Providing) over the Web; link with mobile technologies for retrieving data, loading maintenance action; workflow module, etc [19].

V. AN INTEGRATED FRAMEWORK

According to the e-maintenance architectural framework of Figure 1 the 'wireless' component provides required ICT infrastructure for supporting various activities of the other architectural components. This study extends the model in Figure 1 by providing further insights into the ICT support of the e-maintenance. The argument raised in the study is that the identified ICT components when combined with the e-maintenance architectural framework of Figure 1 together provide an integrated framework for e-maintenance that can be used by today's highly information-intensive manufacturing and service organizations for managing their e-maintenance processes.

VI. SUMMARY AND CONCLUSION

This paper reviewed current literatures in the areas of ICT and Manufacturing/Maintenance in order to provide further insights into the specific ICT requirements of the e-maintenance process. A recent e-maintenance framework, representing the latest effort in the field, was selected on the basis of its relevance to the manufacturing as well as its explicit notion of ICT support as an enabler of e-maintenance activities. The study expanded the ICT component of the framework by investigating current ICT components technologies as a supplement to the existing model, hence the name 'integrated framework'; a framework that integrates a recent e-maintenance framework with specific ICT component technologies.

In future the author intends to extend the current study by providing taxonomy for ICT support of e-maintenance activities and apply the framework to manufacturing industries with the aim of evaluating its suitability to various industries as a step towards developing a generic integrated e-maintenance framework.

REFERENCES

[1] Kajko-Mattsson, M., Karim, R. and Mirijamdotter, A. "Fundamentals of the eMaintenance Concept". *Proceedings of The 1st international workshop and congress on eMaintenance.* 2010, Luleå, Sweden.

[2] Fuchino, T., Shimada, Y., Miyazawa, M. and Naka, Y. "Business process model for knowledge management in plant maintenance". *In:* Bertrand, B. and Xavier, J. (eds.) *Computer Aided Chemical Engineering.* Elsevier. 2008, pp. 101-102.

[3] Marquez, A. C., De Leon, P. M., Fernandez, J. F. G., Marquez, C. P. and Campos, M. L. "The maintenance management framework: A practical view to maintenance management". *Journal of Quality in Maintenance Engineering,*2009; 15(2)**,** pp. 167-178.

[4] Candell, O. *Development of Information Support Solutions for Complex Technical Systems using eMaintenance.* PhD, Luleå University of Technology, 2009.

[5] Karim, R. and Soderholm, P. "Application of information and communication technology for maintenance support information services: Transferring experiences from an eHealth solution in Sweden". *Journal of Quality in Maintenance Engineering,* 2009; 15(1)**,** pp. 78-91.

[6] Han, T. and Yang, B. S. "Development of an e-maintenance system integrating advanced techniques". *Computers in Industry,*2006; 57**,** pp. 569-580.

[7] Sahoo, T. and Parida, A. "Improving Overall Equipment Effectiveness (OEE) of process plant equipments through e-Diagnostics". *Proceedings of The 1st international workshop and congress on eMaintenance.2010,* Luleå, Sweden.

[8] Rasoyska, I., Chebel-Morello, B. and Zerhouni, N. "Process of s-maintenance: Decision support system for maintenance intervention". *Proceedings of the 10th IEEE Conference on Emerging Technologies and Factory Automation, ETFA 2005*, pp. 679-686.

[9] Zhao, Z., Wang, F.-L., Jia, M.-X. and Wang, S. "Predictive maintenance policy based on process data". *Chemometrics and Intelligent Laboratory Systems,* 2010; 103(2)**,** pp. 137-143.

[10] Muller, A., Crespo Marquez, A. and Iung, B. "On the concept of e-maintenance: Review and current research". *Reliability Engineering & System Safety*, 2008; 93(8)**,** pp. 1165-1187.

[11] Muller, A., Suhner, M.-C. and Iung, B. "Formalisation of a new prognosis model for supporting proactive maintenance implementation on industrial system". *Reliability Engineering & System Safety,* 2008; 93(2)**,** 234-253.

[12] Kenne, J. P. and Nkeungoue, L. J. "Simultaneous control of production, preventive and corrective maintenance rates of a failure-prone manufacturing system". *Applied Numerical Mathematics,* 2008; 58(2)**,** pp. 180-194.

[13] De Lucia, A., Pompella, E. and Stefanucci, S. 2005. "Assessing effort estimation models for corrective maintenance through empirical studies". *Information and Software Technology,* 2005; 47(1)**,** pp. 3-15.

[14] Iung, B., Levrat, E., Marquez, A. C. and Erbe, H. "Conceptual framework for e-Maintenance: Illustration by e-Maintenance

 ISBN: 978-1-61208-181-6

technologies and platforms". *Annual Reviews in Control*, 2009; 33(2), 220-229.

[15] Emmanouilidis, C., Liyanage, J. P. and Jantunen, E." Mobile solutions for engineering asset and maintenance management". *Journal of Quality in Maintenance Engineering*, 2009; 15(1), pp. 92-105.

[16] Zhang, Y., Gu, Y., Vlatkovic, V. and Wang, X. "Progress of smart sensor and smart sensor networks". Proceedings of the Fifth World Congress on Intelligent Control and Automation, WCICA 2004(4). pp. 3600-3606.

[17] Voisin, A., Levrat, E., Cocheteux, P. and Iung, B. "Generic prognosis model for proactive maintenance decision support: Application to pre-industrial e-maintenance test bed". *Journal of Intelligent Manufacturing*, 2010; 21(2), pp. 177-193.

[18] Campos, J., Jantunen, E. and Prakash, O. "A web and mobile device architecture for mobile e-maintenance". *International Journal of Advanced Manufacturing Technology*,2009; 45(1-2),pp. 71-80.

[19] Tsang, A. H. C. "Strategic dimensions of maintenance management". *Journal of Quality in Maintenance Engineering*, 2002; 8(1), pp. 7-39.

[20] Jardine, A. K. S., Lin, D. and Banjevic, D. "A review on machinery diagnostics and prognostics implementing condition-based maintenance". *Mechanical Systems and Signal Processing*, 2006; 20(7), pp. 1483-1510.

 ISBN: 978-1-61208-181-6

Semi-Automatic Schema Pre-Integration in the Integration of Modeling Language Independent Behavioral Schemata

Peter Bellström[1], Christian Kop[2], Jürgen Vöhringer[3]

[1]Department of Information Systems, Karlstad University, Karlstad, Sweden
e-mail: Peter.Bellstrom@kau.se

[2]Institute for Applied Informatics, Alpen-Adria Universität Klagenfurt, Klagenfurt, Austria
e-mail: chris@ifit.uni-klu.ac.at

[3]econob GmbH, Klagenfurt, Austria
e-mail: juergen.voehringer@econob.com

***Abstract*—In this paper, we address schema pre-integration in the integration of modeling language independent behavioral schemata. In doing so, we propose and present a set of tasks that should be carried out not only to improve and clarify the meaning of a schema, but also to facilitate the resulting time consuming and error-prone phases in the integration process. Due to the complexity of schema pre-integration, domain experts and repositories (e.g. ontologies) are still important sources of knowledge and should therefore be involved during the whole integration process. As its main contribution, the paper offers new tasks to perform in the schema pre-integration process as well as an adjusted and enriched work of previously presented tasks for schema pre-integration.**

Keywords-Schema Integration; Pre-integration; Behavioral schemata; Dynamic Schemata

I. INTRODUCTION

When designing an information system, the designer and domain experts produce a set of conceptual schemata illustrating both the structural (static) and the behavioral (dynamic) aspects. During the development of larger proposed information systems or enterprise models, these models cannot be built at once into one schema. Often, initially different views are generated, which have to be merged afterwards into the proposed overall schema. Due to the fact that it is one information system, and not a set (cf. set of schemata) that is going to be developed, the conceptual schemata need to be integrated. In another integration scenario, separate information systems based on separate schemata already exist. However, due to various reasons (merging of enterprises, the need for consolidating federated databases for information retrieval etc.), these schemata have to be integrated in order to show the whole picture. Whereas the first scenario is often referred to as "view integration", the second one is called "schema integration". Since both integration scenarios are based on schemata (either final schemata or schema parts), we follow the definition given in [1] and will hereafter use the term "schema integration". Schema integration is described by [1] as "the activity of integrating the schemas of existing or proposed databases into a global, unified schema." (p. 323) and is by the same authors divided into four phases: *pre-integration*, *comparison of the schemata*, *conforming the schemata* and, *merging and restructuring*. We extend this definition only in the aspect that we do not only concentrate on databases. Databases are one application domain of the integration process.

The focus of this paper is the first phase: pre-integration. The reason for focusing on pre-integration is its important impact on the integration step. In the literature, it is already shown that this process can influence how efficiently different schemata can be merged.

This holds especially true if we do not focus on a specific modeling language but try to integrate schemata modeled in different languages for the same purpose. For instance, one schema could be modeled in the business process modeling notation (BPMN) [2] whereas another one could be modeled using ARIS event process chains [3]. In [4], it was shown that this can happen if the business process models of two enterprises that from now on will be in a consortium, must be merged.

Hence, our approach is based on previous approaches on schema integration and particularly on schema pre-integration. The approach is novel, since it tries to introduce modeling language independent schemata, which have the same aim and purpose, as input for the integration process. It focuses on the special tasks that must be considered in such a scenario (i.e. transforming the different modeling elements to a common abstract level).

This paper is therefore structured as follows: in section two, we describe the schema integration process as such. We mainly refer to the important work of [1]. In section three, we address related work on pre-integration and in section four the tasks to perform in schema pre-integration are described. Finally, the paper closes with conclusions and an outlook to future work.

II. THE SCHEMA INTEGRATION PROCESS

In [1], the authors divided the integration process into four activities: *pre-integration*, *comparison of the schemata*,

 ISBN: 978-1-61208-181-6

conforming the schemata and *merging and restructuring*. To grasp what schema integration is all about, each of these four activities will now be shortly addressed and described.

A. Pre-integration

In the pre-integration activity, general analyses of the schemata are applied in order to find strategies for how the schemata have to be integrated. This includes the:

- choice of schemata (views) that have to be integrated into a whole schema,
- collection of additional relevant information that are interesting during integration (i.e. assertions or constraints that must hold among the schemata), and
- strategy (policy) for the integration process (i.e. which schema comes first, which schema is integrated with which other schema).

According to the results described in [1], the integration policy can be a binary and an n-ary policy. The binary integration strategy can be further divided into a ladder-procedure or a balanced procedure. The n-ary approaches can be divided into a one-shot integration and iterative strategy. Integration policy is called a ladder, if there are two schemata that are integrated at the beginning. The first intermediate integration result is then integrated with a third schema. The resulting second intermediate integrated schema is then integrated with a fourth schema and so on. In the balanced strategy, all the source schemata are integrated pairwise. The first intermediate integrated schemata are then integrated pairwise (i.e. the integration of schemata behaves like a binary tree). In the one shot approach, all the schemata are integrated into one global schema at once. If more than 2 schemata are integrated and it is not done in a one-shot strategy, then it is called an iterative strategy., The ordering of schemata and intermediate integrated schemata might be important, especially in the iterative and the ladder strategies. In these approaches, as well as in the balanced approach, it might be important which schema is integrated with which other schema. In [1], advantages and problems of the different strategies are discussed.

B. Comparision of the schemata

In this activity, it must be detected if concepts are the same or are different. Conflicts (i.e. concepts are identical or different) can be classified into naming conflicts. For instance are “employee” and “staff” the same concepts or not? Besides naming conflicts, structural conflicts can also appear in the schemata. For instance, the same concept can be modeled as an attribute in one schema and as an entity type in another schema.

C. Conforming the Schemata

In this activity, conflicts are resolved as best as possible and the schema is transformed (i.e. prepared) for the merging activity. The results of this activity are schemata with schema elements that conform (e.g. in the schemata to be integrated, a concept such as “customer” is an entity type in both schemata).

D. Merging and Restructuring

In this activity, the schemata will be integrated into one schema. Besides conflict resolution described in the previous section, it is also necessary to complete the integrated schema. For instance, if in one schema the concept “employee” exists and in another schema the concept “manager” exists, then it might be necessary to introduce a generalization relationship between manager and employee. Several other operations for this activity are introduced in [1].

III. RELATED WORK

The work presented in [1] and [5] showed that only three works explicitly mentioned the pre-integration step. In the following, other integration approaches were published, which focused on several aspects of the integration problem.

A. Integration of Structure

In [6], the authors discovered that integration of structural schemata can be explained with attribute equivalence. They concluded that this is the basic concept throughout schema integration. The integration process starts with an existing (logical) database schema.

Another work on operators for deciding on the similarity or dissimilarity of schema construct was described in [7].

In [8], the author uses logical assertions to define which constructs of two conceptual models are equivalent. On the basis of these defined assertions, he proposed a method for the automatic integration of the two schemata.

The work presented in [9] concentrates on the automatic detection of naming conflicts.

Algorithms for structural schema integrations are introduced in [10].

Also in the approach presented in [11], the aim is to integrate existing database schemata. For this purpose, they enrich the schema semantically.

An object oriented framework for the integration of heterogeneous databases is presented in [12].

In [13] the authors discuss the impact of linguistic knowledge for the integration step. This approach is based on the well-established assumption that relationships are expressed by a verb. From a linguistic point of view, verbs always have a semantic structure in which nouns play a certain semantic role. For instance if a person buys something, then person has the semantic role of an actor. This knowledge can be used during the merging process (e.g. comparing only actors if two relationships are named with the same verb).

In [14] the authors introduce a black board architecture for schema integration of existing databases. A black board architecture system supports the sharing of knowledge from different knowledge sources. These knowledge sources are the designers and end users who feed the system with their knowledge.

The impact of similarity measures for schema matching and data integration is discussed in [15].

 ISBN: 978-1-61208-181-6

B. Integration of Behavior and other Aspects

Up to now, the described mentioned approaches were developed to integrate structural aspects (i.e. information needed for database design).

Integration of behavioral aspects in object oriented models is mentioned in the work of [16][17]. They describe the integration of state charts of an object type (e.g. life cycle of a book in a library [17]).

In [18], the authors present another interesting work of integration of dynamic object oriented models. Their work is based on the formalization of state chart constructs.

In [19], the authors provide a roadmap for behavior-based integration. They propose a meta class framework on which integration should be based.

An overview of business process integration is given in [20].

OWL-S ontologies are proposed as a support in a method for business process integration [21]. The business process models are firstly transformed into OWL-S models, which are then integrated.

In [22], the authors describe the integration of use cases. The authors exploit information of modular petri nets, which describe these use cases.

Finally requirement statements on behavior are integrated using the behavior tree approach [23]

C. Pre-Integration

Among these above-listed approaches, pre-integration was either explicitly mentioned or, according to the needed input for the integration step, it could be concluded which kind of tasks are necessary in the pre-integration step.

Because the research work of both [6] and [14] used sources that were relational schemata, these sources are firstly brought into a canonical form (i.e. a specific conceptual model). This can be seen as a pre-integration step. In [14], the authors also explicitly mention the pre-integration step. In their work, schema translation and pre-integration are separate steps. Their pre-integration is about making a decision about the policy used to integrate the schemata (i.e. binary, ladder or n-ary strategy).

The need for defining assertions and constraints as mentioned in [6] and in [8] might also be treated as part of a pre-integration step, since the schemata are prepared in order to make integration easier. The step denoted as “assertion specification” or “schema integration assertion” in [6] and in [8] respectively, is subsumed as the schematic inter-schema relation integration in the work of [14]. A similar policy of pre-integration can be found in [11]. Their first step of integration is called “semantic enrichment phase”. This step contains a knowledge acquisition step and a schema conversion step. In the knowledge acquisition, the schemata are analyzed to discover semantics and implicit restrictions (e.g. keys, attribute dependencies). Afterwards, everything is also converted to a canonical form. In their work, this canonical form is a BLOOM schema.

The methodology presented in [13] implies that there are linguistic knowledge bases (i.e. lexicons) that can be used. Furthermore, elements of an entity relationship model must be extended with this extra information. Hence, the use of the lexicons and the generation of extra information are part of a pre-integration step. The outcome of this step is then the input for the subsequent steps.

The work of [16] and [17] has some sort of pre-integration work as well. The first phase is called "integration in the large", whereas the second phase is called "integration in the small". Integration in the large can be seen as a kind of pre-integration, because relationships between state charts are built manually. These relationships express the "dependencies" between state charts which are considered for integration. These relationships support the search for an integration plan in order to reduce expensive integration operations. Integration according to this plan is done in the second phase of their approach (integration in the small).

In the work presented in [19], a pre-integration step is briefly discussed. The pre-integration starts with the definition of the behavior and structure of the meta classes for a domain, which can then be specialized for a certain application.

In the work presented in [21], on an ontology based method for business integration, OWL-S can be seen as the canonical form within a pre-integration step.

The modular petri nets used in [22] can be interpreted as the canonical form to describe the behavior within use cases, which support the integration process.

The behavior tree model is the canonical form for integration of requirements in [23].

As shown in section II and section III, pre-integration is only partially in focus. In this paper, we therefore focus and highlight the pre-integration phase and present new tasks as well as adjusted and enriched tasks that should be used in the pre-integration phase. In the long run, these tasks should not only reduce the time needed for integration but also reduce the risk of errors occuring.

IV. PRE-INTEGRATION IN SCHEMA INTEGRATION

Pre-integration is the first step that should be carried out when conducting schema integration. However, studying related work within the research field shows that this phase has often been overlooked [24], and in some of the early methods, it was not even mentioned [1].

In the semi-automatic method for the integration of modeling language independent behavioral schemata that we are currently researching, we propose conducting the following tasks in the pre-integration step: a) *translate the schemata into one modeling language*, b) *schema constituent name adaption*, c) *schema constituent disambiguation*, d) *standardization of the abstraction level*, e) *recognition and resolution of intra-schema conflicts*, f) *introduction of missing constituents*, g) *selecting the integration strategy*, and h) *selecting the order of integration*.

It should be noted that similar phases have been mentioned for the integration of structural schemata (see for instance [25]). However, in [25], the authors researched the integration of structural schemata, whereas we research the integration of behavioral schemata. It should also be noted that not all tasks were addressed in [25] and if the task was addressed it needs to be adjusted to fit integration of

 ISBN: 978-1-61208-181-6

behavioral schemata, as will be shown and exemplified in this section. Most of the described tasks can be atomized or at least partly atomized contributing to a semi-automatic approach to schema pre-integration.

A. Translate the schemata into one modeling language

The first task to perform in schema pre-integration is to translate all schemata into one modeling language. In [24], this task was called canonization. Choosing the right canonical model for the current project is also emphasized in [26] in which the author focuses on schema translation in federated information systems stating that "[…] the canonical data model must have an expressiveness which is equal to or greater than that of any of the native models in the federation. […] a canonical data model should contain as few basic constructs as possible." (p. 15). In our method, this means translating all schema elements to conditions (pre- and post) and process types, which are the minimal modeling constituents for describing and modeling the behavior of an information system [27]. In [4], this was demonstrated using a small library system prepared for integration. In [4], it was also emphasized that during this transformation, it is important that all labels of the original schema elements are represented in the process type e.g., *Reserve Educational Book* becomes *Customer Reserves Educational Book*, adding the actor into the process type label. This task can often be partly atomized, since tools exist that can aid in the process of translating a schema from one modeling language to another. However, the schema produced by these tools should be viewed as an intermediate schema, since these most likely need to be manually adjusted to meet all rules of the chosen modeling language. In the end, it is still the domain expert that has the domain knowledge and therefore can decide how the processes and states should be combined and described.

B. Schema constituent name adaption

The second task to perform in schema pre-integration is to adapt schema constituent names to specific naming rules. The names of constituents are very important and if they are put together, they should reflect the meaning of either a condition or a process type. However, to facilitate semi-automatic schema integration, they should not only be readable for humans, e.g. domain experts and designers, but also be readable to a computerized application. This means that a formal language is not useful but neither is natural language since it is ambiguous. We therefore use naming guidelines to adjust the language used in the schemata and thereby end up with a controlled subset of natural language. In [27] and [28], this was first mentioned as an important guideline, called *standardization of concept notions*.

Three examples on how to use this task and when it is applicable are as follows:

- name static concepts in ***singular*** (e.g. *Books → Book*)
- name process types with the ***verb + noun*** rule (e.g. order → order Book)
- name conditions with the ***noun + verbal principal*** rule (e.g. *reserved → Book reserved).*

Since we assume that integration of structural schemata is conducted before integration of behavioral schemata, the integrated structural schema can also be used as an information repository; it can even be used as a template on how to adjust constituent names.

Finally, similar and complementary approaches are given in [28] and [29], where the authors described a controlled langauge approach for OWL verbalization [28] and schema constituent adjustment [29].

C. Schema constitutent disambiguation

The third task to perform in schema pre-integration is schema constituent disambiguation. In this task, we add descriptions and definitions of process types and conditions. This task could either be done manually by domain experts or automatic suggestions could be generated using domain ontologies or general lexicons such as Wordnet [30]. However, to get a good and reliable result from this step, it is important that prior to this task schema constituent name adaption has been conducted. Since we also assume that integration of structural schemata is done prior to integration of behavioral schemata, the results from that task could also be used within schema constituent disambiguation. For instance, having conducted schema element disambiguation for static schemata [25], we have already collected definitions and explanations of important structural concepts that could be reused. The integrated structural schema also indicates which structural concepts need to be processed and given conditions (see also schema constituent name adaption). This task is rather complex and often we need to split sentences or sequences of words into single words and from that move on with the disambiguation task [28].

D. Standardization of the abstraction level

The fourth task to perform in pre-integration is standardization of the abstraction level. In [16] and [17], the authors address integration of state charts, mentioning the problem of *state overlapping*, meaning one state in a specific state chart corresponds partially to a specific state in an another state chart. In [27], it was mentioned that similar problems were identified in the integration of behavioral schemata, where they were called *process type overlapping* and *condition overlapping*.

In our method, we address the overlap problem by trying to standardize the abstraction level in each behavioral schema. In doing so, we agree with [28] who also address the problem of different abstraction levels, stating that the schemata should be detailed without addressing implementation issues and each modeling element should be atomic. If for instance a process type is recognized as not being atomic, the process type needs to be analyzed and modified to fulfill this criterion.

E. Recognition and resolution of intra-schema conflicts

The fifth task to perform in schema pre-integration is recognition and resolution of intra-schema conflicts. This means analyzing one single source schema aiming to recognize conflicts (similarities and differences) within the schema. This is an important task since oftentimes two

 ISBN: 978-1-61208-181-6

process types or conditions are named the same but the actual meaning is very different, or that one process type or condition is given different names but the actual meaning is the same. In other words, in this task we look for potential homonym and synonym conflicts. To do so, we not only analyze and compare the name of the constituents, but also the neighborhoods (surrounding). Comparing the neighborhood of the constituent has also been addressed in relation to integration of structural schemata. In [31] and [32] for example, the authors use neighborhood comparison as a matching strategy during semi-automatic integration of modeling language independent structural schemata. Similar techniques are also used in the DIKE approach [33] and the GeRoMeSuite [34]. However, our approach is placed much closer to the work presented in [31] and [32], due to the focus on modeling language independent integration.

F. Introduction of missing constituents

The sixth task to perform in schema pre-integration is the introduction of missing constituents. More precisely, this means introducing missing process types, conditions or connections between constituents. During several of the described pre-integration tasks, e.g. translate the schemata into one modeling language and standardization of the abstraction level, the domain expert and designer might identify holes: some constituents are missing in the schema that is currently being prepared for integration. In this task, the process types, conditions and connections between them that are identified by the domain experts and designers are added manually. If a domain ontology and/or taxonomy are available, these can be used to recognize a missing connection between two constituents. The possibility of using a behavioral taxonomy to enrich behavioral schemata was also addressed by [19]. However, their approach focused on the object oriented modeling paradigm, while we instead research modeling language independent integration of behavioral schemata.

G. Selecting the integration strategy

The seventh task to perform in schema pre-integration is selecting the integration strategy. In this task, the order of integration is decided. The integration order can be divided into binary and n-ary integration. Binary integration can further be divided into binary ladder and binary balanced and n-ary integration into n-ary one-shot and n-ary iterative [1].

In our approach, we have decided to use binary ladder, meaning two schemata are always integrated (see section II A). Using binary ladder is preferred since the complexity is reduced due to only processing two schemata within each iteration and we can also in a semi-automatic way decide upon a first suggestion of the order of integration.

During the 1980's, this task was also the main task researched for pre-integration in the integration of structural schemata (e.g. [5][35]).

H. Selecting the order of integration

The last task to perform in schema pre-integration is selecting the order of integration. Having decided to use binary ladder (see the former task), this task should result in a decision regarding the specific integration order, meaning which schemata should first be integrated and so on. To decide upon the specific integration order. we not only analyze and compare the pre-condition(s) and the post-condition(s) of each schema, but we also analyze the process type descriptions. By analyzing and comparing the conditions, we mean the first and last conditions of each schema. For instance, the first condition of schema one is the pre-condition *Book Not Reserved* and the last post-condition of the same schema is *Book Reserved*. This means that conditions within the schema are in this step viewed as a black box. In other words, in this task we are looking for conditions that might be the same. For instance, two disjoint schemata have the same post-conditions, a post-condition in one schema is a pre-condition in another (consecutive), two parallel schemata have the same pre- and/or post-conditions, two alternative schemata have the same pre- and/or post-conditions and finally two schemata are viewed as equivalent having the same pre-and post conditions. For a more detailed discussion and description of schema relationship types, please see [16][17], who research the integration of state charts and [28], who research integration of Klagenfurt Conceptual Pre-Design Models [36].

To complement the analysis and comparison of conditions, we also count the occurrences of the most important terms used within the schemata. This should be done since the resulting figures could aid in the process of deciding the order or integration. For instance, if schema one describes the process of storing an order and schema two describes the process of updating an already stored order, counting the number of "order" should most likely indicate that these two schemata should be integrated in one iteration.

Selecting the order of integration is also facilitated if the task of standardization of the abstraction level has already been carried out. Additionally, the integrated structural schemata can be used as a knowledge source, since in it the static data are defined and described.

V. CONCLUSION AND FUTURE WORK

In this paper, we have addressed schema pre-integration in the integration of language independent behavioral schemata. In doing so, we have presented and described a set of tasks that are important to carry out to facilitate the resulting time consuming and error-prone integration process.

In this paper, it has been shown that in all presented tasks, some type of electronic knowledge source could and should be used to assist the domain experts and designers, contributing to a semi-automatic approach to schema pre-processing.

In future research, we will continue our work on developing a semi-automatic method for the integration of modeling language independent behavioral schemata. In doing so, we will amongst other things research how to use knowledge sources, such as ontologies, taxonomies, dictionaries and lexicons, in the entire integration process.

 ISBN: 978-1-61208-181-6

REFERENCES

[1] C. Batini, M. Lenzerini, and S.B. Navathe, "A Compartive Analysis of Methodologies for Database Schema Integration," ACM Computing Surveys, vol. 18(4), 1986, pp. 323-364.

[2] BPMN – Business Process Modeling Notation, [Electronic], Available: http://www.bpmn.org/ [20111108].

[3] A.W. Scheer, ARIS – Business Process Frameworks, Heidelberg: Springer, 1999.

[4] P. Bellström, J. Vöhringer, and C. Kop, "Toward Modeling Language Independent Integration of Dynamic Schemata," Information Systems Development Toward a Service Provision Socity, Heidelberg: Springer, 2009, pp. 21-29.

[5] C. Batini and M.Lenzerini, "A Methodology for Data Schema Integration in the Entity-Relationship Model," IEEE Transactions on Software Engineering, 10 (6), 1984, pp. 650-664.

[6] J.A. Larson, S.B. Navathe, and R. Emasri, "A Theory of Attribute Equivalence in Databases with Application to Schema Integration," Transactions on Software Engineering, Vol. 15 (4), 449-463.

[7] A. Savasere, A. Sheth, and S. Gala, "On Applying Classification to Schema Integration," Proc. First International Workshop on Interoperability in Multidatabase Systems (IMS'91), IEEE Press, 1991, pp. 258-261.

[8] P. Johannesson, "A Logical Basis for Schema Integration," Third International Workshop on Research Issues on Data Engineering (RIDE-IMS'93), IEEE Press, 1993, pp. 86-95.

[9] H. K. Bhargava and R.M. Beyer, "Automated Detection of Naming Conflicts in Schema Integration: Experiments with Quiddities," Proc. 25th Hawaii International Conference on System Sciences, IEEE Press, 1992, pp. 300-310.

[10] J. Geller, A. Mehta, Y. Perl, E. Neuhold, and A. Sheth, "Algorithms for Structural Schema Integration," Proc. Second International Conference on Systems Integration (ICSI'92), IEEEP Press, 1992, pp. 604-614.

[11] M. García-Solaco, F. Saltor, and M. Castellanos, "A Structure Based Schema Integration Methodology," Proc. Eleventh International Conference on Data Engineering, IEEE Press, 1995, pp. 505-512.

[12] H. Dai, "An Object-Oriented Approach to Schema Integration and Data Mining in Multiple Databases," Proc. Technology of Object-Oriented Languages (TOOLS), IEEE Press, 1997, pp. 294-303.

[13] E. Métais, Z Kedad, I. Comyn-Wattiau, and M. Bouzeghoub, "Using Linguistic Knowledge in View Integration: Toward a Third Generation of Tools," Data & Knowledge Engineering 23(1), 1997, 59-78.

[14] S. Ram and V. Ramesh, "A Blackboard-Based Cooperative System for Schema Integration," IEEE Expert, June 1995, 56-62.

[15] A. Gal, "Interpreting Similarity Measures: Bridging the Gap between Schema Matching and Data Integration," Data Engineering Workshop of ICDEW 2008, IEEE Press, 2008, pp. 278-285.

[16] H. Frank and J. Eder, "Towards an Automatic Integration of Statecharts," International Conference on Conceptual Modeling (ER 1999), Heidelberg: Springer, 1999, pp. 430-444.

[17] H. Frank and J. Eder, "Integration of Behavioral Models," Proc. ER'97 Workshop on Behavioral Models and Design Transformations: Issues and Opportunities in Conceptual Modeling, [Electronic], Available: http://osm7.cs.byu.edu/ER97/workshop4/fe.html [20111108], 1997.

[18] B.H.C. Cheng. and E. Y. Wang, "Formalizing and Integrating the Dynamic Model for Object Oriented Modeling," IEEE Transactions on Software Engineering, Vol 28 (8), August 2002, 747-762.

[19] M. Stumptner, M. Schrefl, and G. Grossmann, "On the Road to Behavior-Based Integration," Proc. 1st APCCM Conference, 2004, pp. 15-22.

[20] A. Raut, "Enterprise Business Process Integration," Conference on Convergent Technologies for Asia-Pacific Region, IEEE Press, 2003, pp. 1549-1553.

[21] S. Fan, L. Zhang, and Z. Sung, "An Ontology Based Method for Business Process Integration," International Conference on Interoperability for Enterprise Software and Applications in China, IEEE Press, 2008, pp. 135-139.

[22] W. J. Lee, S. D. Cha, and Y. R. Kwon, "Integration and Analysis of Use Cases Using Modular Petri Nets in Requirements Engineering," IEEE Transaction of Software Engineering, Vol. 24 (12), December 1998, 1115-1130.

[23] K. Winter, I.J. Hayes, and R. Colvin, "Integrating Requirements: The Behavior Tree Philosophy," 8th IEEE International Conference on Conference on Software Engineering and Formal Methods (SEFM), IEEE Press, 2010, pp. 41-50.

[24] W. Song, Schema Integration – Principles, Methods, and Applications, Dissertation, Stockholm University, 1995.

[25] P. Bellström and J. Vöhringer, "A Semi-Automatic Method for Matching Schema Elements in the Integration of Structural Pre-Design Schemata," unpublished.

[26] P. Johannesson, Schema Integration, Schema Translation, and Interoperability in Federated Information Systems, Dissertation, Stockholm University, 1993.

[27] P. Bellström, J. Vöhringer, and C. Kop, "Guidelines for Modeling Language Independent Integration of Dynamic Schemata," Proc. IASTED International Conference on Software Engineering, 2008, pp. 112-117.

[28] J. Vöhringer, Schema Integration on the Predesign Level, Dissertation, Alpen-Adria-Universität Klagenfurt, 2010.

[29] G. Fliedl, C. Kop, and J. Vöhringer, "From OWL class and property labels to human understandable natural language," Proc. NLDB'07, 2007, pp. 156-167.

[30] G. A. Miller, "WordNet: A Lexical Database for English," Communications of the ACM, 38 (11), 1995, 39-41.

[31] P. Bellström, and J. Vöhringer, "Towards the Automation of Modeling Language Independent Schema Integration," International Conference on Information, Process, and Knowledge Management (eKNOW 2009), IEEE computer society, 2009, pp. 110-115.

[32] P. Bellström and J. Vöhringer, "A Three-Tier Matching Strategy for Predesign Schema Elements," The Third International Conference on Information, Process, and Knowledge Management (eKNOW 2011), 2011, pp. 24-29.

[33] L. Palopoli, G. Terracina, and D. Ursino, "DIKE; A System Supporting the Semi-Automatic Construction of Cooperative Information Systems From Heterogeneous Databases," Software–Practice and Experiences, vol. 33, 2003, pp. 847-884.

[34] D. Kensche, C. Quix, X. Li, and Y. Li, "GeRoMeSuite: A System for Holistic Generic Model Mangement," Proc. 33rd International Conference on Very Large Data, 2007, pp. 1322-1325.

[35] S. B. Navathe and S.U. Gadgil, "A Methodology for View Integration in Logical Database Design," Proc. Eighth International Conference on Large Data Bases, Morgan Kaufmann, 1982, pp. 142–164.

[36] G. Fliedl, C. Kop, H.C. Mayr, W. Mayerthaler, and C. Winkler, "Linguistically based requirements engineering – The NIBA project," Data & Knowledge Engineering, 35, 2000,111-120.

 ISBN: 978-1-61208-181-6

Automatic Keyphrase Extraction: A Comparison of Methods

Richard Hussey, Shirley Williams, Richard Mitchell
School of Systems Engineering
University of Reading
Reading, United Kingdom
{r.j.hussey, shirley.williams, r.j.mitchell}@reading.ac.uk

***Abstract*—There are many published methods available for creating keyphrases for documents. Previous work in the field has shown that in a significant proportion of cases author selected keyphrases are not appropriate for the document they accompany. This requires the use of such automated methods to improve the use of keyphrases. Often the keyphrases are not updated when the focus of a paper changes or include keyphrases that are more classificatory than explanatory. The published methods are all evaluated using different corpora, typically one relevant to their field of study. 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 six corpora. The methods chosen were term frequency, inverse document frequency, the C-Value, the NC-Value, and a synonym based approach. These methods were compared to evaluate performance and quality of results, and to provide a future benchmark. It is shown that, with the comparison metric used for this study Term Frequency and Inverse Document Frequency were the best algorithms, with the synonym based approach following them. Further work in the area is required to determine an appropriate (or more appropriate) comparison metric.**

Keywords- Term Frequency, Inverse Document Frequency, C-Value, NC-Value, Synonyms, Comparisons, Automated Keyphrase Extraction, Document Classification.

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.

The initial decision to make this comparison study stemmed from earlier work [1] that had shown a tendency on the part of authors to use corpora from their discipline area. For example, those of a medical background used medical corpora such as the PubMed Central database, while those in literature might use the Journal on Applied Linguistics. This made the task of comparing the effectiveness of one method to another more complex, as there was no common thread or baseline.

This comparison study builds on a pilot study [1] that showed for a small number of algorithms and corpora, the Term Frequency method was the best. Therefore, this paper sets out to compare the outputs of five systems on the same six corpora. The methods chosen were all in the field of AKE. The C-Value [2] (and its follow-on the NC-Value [2]) demonstrated a series of linguistic filters for determining what phrases should be considered, and uses a ranking metric based on sub-strings. Hussey et al [3] showed that using a thesaurus to group synonyms into keyphrases could be used to improve the results from analysing the document for common themes.

A. Background

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 [4] (using the Technorati blog [5] as their source document) that a small number of keywords and keyphrases tend to be used (or reused) frequently, while a much larger number are idiosyncratic and demonstrate a low frequency as they are too specific to be reused even by the same author (examples include). Examples of reused phrases included "politics" and "shopping" [5], while "insomnia due to quail wailing" and "streetball china" [5] were among the examples of the idiosyncratic phrases. 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 automatic keyphrase extraction (AKE) is to select the small collection of relevant words that can be used to describe or categorise the document. The process of AKE is discussed by Frank et al. [6]. AKE is characterised by using phrases from the source document (or a reference document).

The rest of the paper comprises a review of the algorithms (Section II), the implementation (Section III), and results gained (Section IV), a discussion of the outcomes (Section V), conclusions, and future work (Section VI).

 ISBN: 978-1-61208-181-6

II. REVIEW

In this section, relevant methods and the associated results are discussed. The term frequency and term frequency inverse document frequency methods are pure statistical methods, and their generic use is discussed first.

A. Term Frequency and Inverse Document Frequency

The term frequency is simply the number of times a given term (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'.

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 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 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 inverse document frequency 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").

B. C-Value

The C-Value algorithm [2] 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, where as Filter 3 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.

$$Cvalue(a) = \begin{cases} \log_2|a| \cdot f(a) & a \text{ is not } a \text{ sub} - string \\ \log_2|a| \cdot \left(f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b) \right) & else \end{cases} \quad (5)$$

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 best phrases are selected for use as keyphrases.

Frantziy et al. 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%.

C. NC-Value

The NC-Value [2] extends the C-Value algorithm by using the words adjacent to of 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

 ISBN: 978-1-61208-181-6

- n is the total number of phrases

This is then fed into Equation 7, the NC-Value. (7)

$$NCvalue(a) = 0.8Cvalue(a) + 0.2 \sum_{b \in C_a} f_a(b) weight(b)$$

The values of 0.8 and 0.2 used were arrived at following experimentation by Frantziy et al. [2], and therefore may only be applicable to the medical corpora they used.

Frantziy et al. compared the C-Value and NC-Value using the previous defined the Recall and Precision metrics (see Section III.D). 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.

D. Roget and Synoynms

The synonym algorithm [3] 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" [7] and therefore both may have been present as separate keyphrases. Therefore, an additional clustering element is added to group these keyphrases into their synonym groups to prevent a single, popular, concept from dominating. This simply involves applying the original Equation 8 to the keyphrases.

III. IMPLEMENTATION

The basis of the work presented here is the examination of a document with the intention of generating keyphrases from it. The implementation details that follow explain the algorithms of the methods compared. Where more than one method or set of configurations was presented, the chosen settings implemented were those that the authors of those papers found produced the best output.

The different algorithms were tested against six corpora. For this study of the algorithms, it was decided that the corpora would be restricted to academic papers, which for the majority case are submitted with keywords against which the results can be tested.

Five corpora were taken from the Academics Conferences International (ACI) e-journals [8], each corpus on a different subject area: *Business Research Methods* (EJBRM), *E-Government* (EJEG), *E-Learning* (EJEL), *Information Systems Evaluation* (EJISE), and *Knowledge Management* (EJKM). The sixth corpus was taken from PubMed Central (PMC) [9], an archive of biomedical and life science journal papers. The thesaurus used was Roget's "*Thesaurus of English Words and Phrases*" [7] and the Porter Stemming algorithm [10] was used to stem the unigrams.

A. Chance Study

For the chance study, the words from the source document were split into a list of individual words. From this list, a start point was chosen at random and a number of contiguous words were strung together to form a keyphrase. The maximum length of the keyphrase was set at $n = 7$ as this was the longest phrase in the reference document, Roget's Thesaurus [7]. The algorithm used was:

- Randomly select a word in the source document to act as a starting point.
- After each word is added, generate a random number less than or equal to n. If this number is greater than the number of words already in the phrase, add another word.
- Repeat until r keyphrases have been produced (in this study, r was chosen to be 5).

B. Term Frequency

For the term frequency (TF) study, the source document was split into a list of single words and a count of the number of times each appeared in the source document occurred, and then normalised to prevent bias towards longer document. The results were then ranked in size order, and the top five taken as the results. The algorithm for this was:

- Count the occurrence of each word in the document
- Divide the count by the total number of words
- Sort the list by frequency and output the top r ranked items (in this study, r was chosen to be 5)

C. Term Frequency – Inverse Document Frequency

The inverse document frequency (IDF) extends the TF algorithm in an attempt to automatically remove from consideration phrases that are not important as keyphrases because they are common to the whole corpus.

- Count the occurrence of each word in the document
- Take the logarithm of the number of documents in the corpus divided by the number of documents containing that word
- Multiple the two values
- Sort the list by frequency and output the top r ranked items (in this study, r was chosen to be 5)

D. The C-Value

For the C-Value, the document was first filtered through a series of linguistic filters as set out in Section II.B. Phrases

 ISBN: 978-1-61208-181-6

that pass the filter are then categorised as either being stand-alone or as being a sub-phrase of another. Those that are stand-alone are assigned a C-Value based on the base-2 log of their number of words. Those that are sub-phrases are given a more complex C-Value as explained in Section II.B and in the algorithm given below:

- Sift phrases through linguistic filter
- For each valid phrase, take the log (base 2) of the number of words in it, and check to see if it is nested inside another valid phrase
    - If it is not then multiply by the number of times it occurs
    - If it is nested, then multiply by the number of time it occurs minus the sum of the number of times all the longer phrases occur divided by the number of those phrases
- Sort the list by frequency and output the top *r* ranked items (in this study, *r* was chosen to be 5).

E. The NC-Value

The NC-Value extends the C-Value by taking the output from that system and adding on a contextual weight. The weight takes into account the words that surround the candidate phrase. The algorithm for this is:

- Sift phrases through linguistic filter
- For each valid phrase, take the log (base 2) of the number of words in it, and check to see if it is nested inside another valid phrase
    - If it is not then multiply by the number of times it occurs
    - If it is nested, then multiply by the number of time it occurs minus the sum of the number of times all the longer phrases occur divided by the number of those phrases
- Multiply this by 0.8 and then add 0.2 multiplied by the context weight for the phrase
    - The weight is the number of phrases that the context word occurs adjacent to divided by the total number of phrases
- Sort the list by frequency and output the top *r* ranked items (in this study, *r* was chosen to be 5).

F. Roget and Synonyms

The synonym study grouped words in the document together by also including counts of their synonyms in the process. This, however, produced a set of keyphrases that tended to be synonyms of each other, so the synonym study further extended the method to involve a second round of clustering to try to prevent a single, popular concept from dominating. For example, the words "acquisition" or "taking" can both mean "recovery" [7] and therefore both may have been present as separate keyphrases.

- For each *n*-gram in the thesaurus, compare the *n*-gram to the associated synonyms
- For each synonym that matches, add the word to a list, and increase its frequency value by the value of the *n*-gram divided by the number of associated synonyms
- Then, for each Key entry in the thesaurus check to see if the frequency is equal to the highest frequency value in the found in the preceding step.
- For each synonym entry associated with the Key, add the synonym to a second list of words and increase its value by one.
- Sort the list by frequency and output the top *r* ranked items (in this study, *r* was chosen to be 5).

IV. RESULTS

The following section sets out the results of the five algorithms studied. For each of the papers analysed in the corpora, the authors (for the most part) supplied an accompanying list of keyphrases to summarise the content. The results of the algorithms were automatically evaluated by comparing them to the author keyphrases. Where a paper did not have author-supplied keywords, it was automatically excluded from the study and the results.

A match was recorded for a paper if at least one of the output keyphrases matched one of the author keyphrases. However, a naïve text matching approach was used for this initial study. The approach would match any two strings if they were either equivalent or a sub-string of the other, e.g. "know" and "knowledge" would be considered a match.

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'. The 'Increase' column, where it occurs, is the numerical value by which the percentage differs from the chance results – i.e. if the match percentage was 1% in the chance study and 10% in the TF study, then that would be an increase of nine.

The results are also summarised in Figure 1 on page 5.

A. Chance Study

The chance results showed almost no keyphrases being produced that matched the authors'. The results can be seen in Table I.

TABLE I. CHANCE RESULTS

Corpus	Papers	Matched	Percentage
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%
PMC	137	1	0.73%
Average			1.44%

B. Term Frequency

Table II shows the results from the term frequency study, and that it performed very well matching on average over 80% of the keyphrases against the authors'.

 ISBN: 978-1-61208-181-6

TABLE II. TF RESULTS

Corpus	Papers	Matched	Percentage	*Increase*
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*
PMC	137	105	76.64%	*75.91*
Average			87.36%	*85.92*

C. Term Frequency – Inverse Document Frequency

The inverse document algorithm showed a drop in performance compared to the simple term frequency results, as shown in Table III.

TABLE III. TF*IDF RESULTS

Corpus	Papers	Matched	Percentage	*Increase*
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*
PMC	137	107	78.10%	*77.37*
Average			69.45%	*68.01*

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. Contrary to expectations based on the original paper [2], the broader the filter the better the results.

TABLE IV. C-VALUE RESULTS

Corpus	Papers	Matched	Percentage	*Increase*
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*
PMC	137	25-31	~21.17%	*~20.44*
Average			~14.47%	*~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	Percentage	*Increase*
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*
PMC	137	0	0.00%	*-0.73*
Average			~0.51%	*~-0.93*

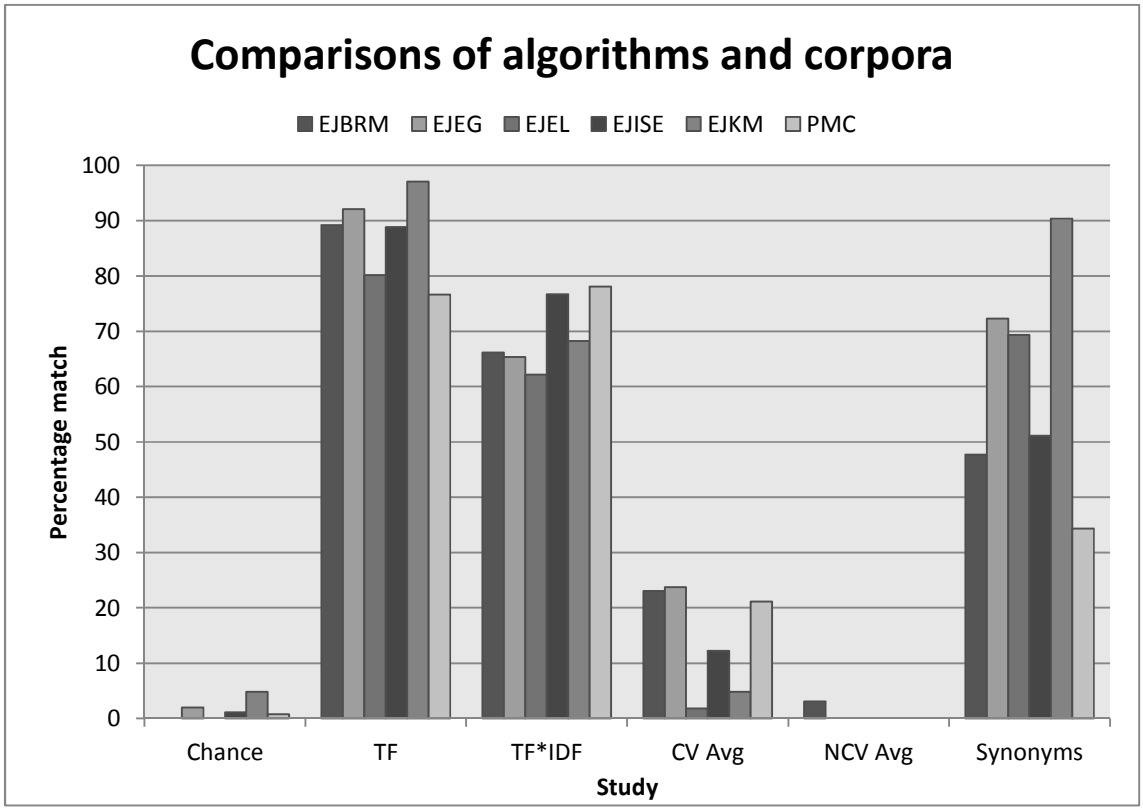


Figure 1

F. Roget and Synonyms

The synonym results show a good improvement over the baseline results (nearly 60% 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. ROGET AND SYNONYM RESULTS

Corpus	Papers	Matched	Percentage	*Increase*
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*
PMC	137	47	34.31%	*33.58*
Average			60.86%	*59.42*

V. DISCUSSION

The results outlined in Section IV above show that the Term Frequency algorithm performed best – the pure statistical method of simply counting how often a word occurred in the document.

However, as has been pointed out in the results section, the matching criteria was naïve. It is perhaps the contention of this paper, but a suggested reason for this is that keyphrases supplied by the paper 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, it is likely that a paper in the e-Journal on Knowledge Management both frequently uses the phrase "knowledge" and has it as an author-assigned keyphrase.

Furthermore, as shown by Sood et al. [4] author-assigned keywords are inappropriate chosen 51.15% of the time. These factors combined suggest that the matching criteria needs to be changed for future work and a Recall/Precision model as used by Frantziy et al. would seem appropriate.

VI. CONCLUSION AND FURTHER WORK

This study has shown that for the naïve comparison method used the results are biased towards phrases that occur most often in the document. However, further studies need to be run with a more standard set of evaluation criteria (such as Recall and Precision) and to be tested on a wider range of corpora – including Reuters-21578 corpus and the remainder of the PMC corpus.

In addition, work needs to be undertaken to validate the outputs of the algorithms by human judges to assess the suitability of both the keyphrases provided by the authors and by the algorithms.

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. 2011. "A Comparison of Methods for Automatic Document Classification", Proceedings of BAAL, The Forty Fourth Annual Meeting of the British Association for Applied Linguistics. Bristol, United Kingdom.

[2] 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.

[3] 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: 5 September 2011]

[4] 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: 13 December 2010]

[5] Technorati. 2006. "Technorati". http://www.technorati.com [Last accessed: 13 December 2010]

[6] 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.

[7] P.M. Roget. 1911. "Roget's Thesaurus of English Words and Phrases (Index)". http://www.gutenberg.org/etext/10681 [Last accessed: 13 December 2010]

[8] Academics Conferences International. 2009. "ACI E-Journals". http://academic-conferences.org/ejournals.htm [Last accessed: 13 December 2010]

[9] PubMed Central. 2011. "PubMed Central Open Access Subset". http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/ [Last accessed: 14 September 2011]

[10] M.F. Porter. 1980. "An algorithm for suffix stripping", Program, 14(3) pp. 130–137.

 ISBN: 978-1-61208-181-6

How to Acquire Scientific Knowledge for University to Industry Knowledge Transfer

Ioana Suciu, Christophe Fernandez, Amadou Ndiaye
Institut National de la Recherche Agronomique,
INRA UMR 927 (Université Bordeaux 1 – CNRS – INRA)
Bordeaux, France
ioana.suciu@bordeaux.inra.fr,
christophe.fernandez@bordeaux.inra.fr,
amadou.ndiaye@bordeaux.inra.fr

Benoit Le Blanc
Institut Polytechnique de Bordeaux, ENSC
Bordeaux, France
benoit.leblanc@ensc.fr

Catriona Raboutet
Real Cognitic
Bordeaux, France
catriona.raboutet@gmail.com

Abstract - **This paper presents a framework dedicated to support the acquisition of useful and usable scientific knowledge that can be transferred to industrialists. It exposes the seven corpora of knowledge (Initial data, Problem, Hypothesis, Tests, Results, Interpretation and Conclusion) necessary both to acquire and to structure relevant scientific knowledge. The framework is used in the domain of design of tailored biscuits with optimized satiety benefit. The acquired scientific knowledge will be represented and next implemented in an electronic knowledge book that is the channel chosen to transfer it to biscuit-making industrialists.**

Keywords-scientific knowledge acquisition; university-to-industry knowledge transfer; knowledge sheets; electronic k-book.

I. INTRODUCTION

Our work is situated in the field of university to industry knowledge transfer (U2I-KT). The objective of the U2I-KT research is to make the knowledge from scientific results available in a format that is directly accessible and exploitable by industrialists. However, scientific knowledge produced by academic research is intended to scientific communities for an academic use. It is commonly made available through scientific publications, which are not easy to be interpreted and used by industrialists. The difficulty comes from the difference that researchers produce scientific knowledge with the aim of understanding problematic phenomena, whereas industrialists demand this knowledge to manufacture their products. The challenge is to reformulate scientific knowledge created in university with respect to the industrial process of product manufacturing. The U2I knowledge reformulation issue will introduce new ways to structure, to access and to handle useful scientific knowledge, all oriented to its final industrial use.

The U2I reformulation program requests a knowledge base (k-base) of relevant and useful scientific knowledge that should be constituted and made available in a form that eases its future treatment (easy access to knowledge should be envisaged, for example). Scientific knowledge should be captured, modelled, structured and represented within the k-base and then prepared for reformulation (re-structured) with respect to a configuration properly shaped for its industrial use. An appropriate-designed tool should be conceived and implemented to easily supply the scientific knowledge formerly reconfigured. It will be exploited to transmit this knowledge to the industrialists in order to be assimilated and used by them. An electronic knowledge book (e-k-book) is the tool chosen here for the U2I-KT. It is a hypermedia electronic document composed by a set of hierarchical concept maps and knowledge sheets (see Section V). The e-k-book has the principal advantage of providing to industrialists direct access and immediate use of knowledge within the k-base. It can be easily handled as being produced on, published through, and readable on a computational form.

The work presented in this paper deals with the acquisition of scientific knowledge. The framework proposed here to *A*cquire *s*cientific *K*nowledge (AsK) is intended to support the k-base constitution and its structuring from a scientific-oriented perspective. Further work will be carried out on scientific knowledge representation, its industry-oriented reformulation and e-k-book implementation.

The paper is structured in six sections: Section I points out the context and the objective of our work. Section II refers to related works on knowledge acquisition and transfer. Section III depicts a structural view over the scientific knowledge. Section IV exposes the AsK framework based on this view. It proposes an aid to acquire and to structure scientific knowledge that will be reformulated. Section V discusses an application of AsK to acquire knowledge in the domain of design of satiety biscuits. Section VI exposes conclusions and future work.

II. BACKGROUND AND RELATED WORK

As cited by [1], knowledge transfer is a complex process, which includes knowledge transmission, absorption and use. Important types, methods and levels for its exploitation are cited in [2]. However, to transfer knowledge one has to acquire and make it available in a way that is suitable to its absorption and use [3]. The frequently used concepts [4] in the context of U2I-KT [5] are present in the knowledge acquisition, too. The absorptive capacity [6] depends on the ability of knowledge use, but also on the capacity to acquire and assimilate it. Prior knowledge and domain similarity [4] facilitate the absorption, but also the acquisition strategies. While a survey of tools and techniques for knowledge acquisition is presented in [7], the closest related work to our research is found in [8]. Here,

 ISBN: 978-1-61208-181-6

knowledge acquisition is "the activity of capturing, structuring and modelling knowledge from any source for the purpose of storing, sharing or implementing knowledge". Thus, the overall objective of the acquisition process is to develop a k-base operational for the future processing. This k-base contains informal, structured, formal or computational know-ledge related to all the relevant and useful aspects. Accordingly to [9], lack of knowledge pertaining to research utilization can inhibit appropriate and effective use. In this background, the originality of our work consists in proposing a framework to acquire and to structure scientific knowledge for the U2I-KT.

III. WHICH STRUCTURE FOR SCIENTIFIC KNOWLEDGE

Scientific knowledge is the key ingredient in the process of U2I-KT and the key object of our study. Published knowledge produced by scientist researchers as a result of their scientific research is evaluated by a scientific community and released through scientific publications. Scientific papers are thus the vehicle of spreading and supplying scientific knowledge. They are regularly structured into five formal parts: (a) Introduction and Literature Review; (b) Material and Method; (c) Results; (d) Discussion; (e) Conclusion and Future Work. Knowledge contained in these publications concerns existing data used for the research and new data provided by this. Inspired by [10], which cites the main steps commonly used in a hypothetical-deductive scientific method, we assume that relevant scientific knowledge encloses data, information and knowledge related to the following corpora: *Initial data*, *Problem, Hypothesis, Tests, Results, Interpretation* and *Conclusion*. These corpora offer a structural perspective on the scientific knowledge. Commonly, *Initial data, Problem and Hypothesis* are found in the (a) section, *Tests* in the (b) section, *Results* in (c), *Interpretation* in (d) and *Conclusion* in the (e) section of a scientific publication.

Initial data includes theories, models, facts, representations, observations, methods or concepts related to a specific domain. It reveals accepted shared knowledge of a scientific community that outlines the prior knowledge needed for the research. *Problem* suggests the hindrance raised from the contradictions between new facts and old ideas. It is formulated by a research main question. *Hypothesis* conveys assumptions formed on the basis of the research problem. *Tests* include *Material* (*Product* and *Equipment* used for experimentation, *Instrument* required to observe or to measure parameters and *Operator* to perform tasks) and *Method* (*Protocol* and *Model* used for investigation, or production). Tests (experimental, simulation- or modelling-oriented) are set up to: i) test existing theories or probe results from the background observation, ii) answer a question or investigate a problem, or iii) test the hypothesis introduced, to support or to disprove them. *Results* present data or interesting observations made when handling data, and report the accurate results of tests. *Interpretation* renders the results explanations, analysis, reflexions and assimilations. *Conclusion* exposes the impact of the new research findings on the current knowledge and the restatement of the knowledge after their integration.

These 7 corpora of knowledge are necessary to acquire and structure relevant scientific knowledge. Certainly, industrialists are in quest of accurate and well interpreted scientific results. Yet they are also interested in data, information and knowledge about essential methods and materials to use, parameters values to adjust, concepts and theories to be familiar with. These will provide them a better understanding of the scientific results and may ease the reproduction of the tests, if needed. Hence, all the available corpora of knowledge have to be collected and handy at the end of the acquisition. Otherwise, lost or unavailable pieces of knowledge can have a negative impact on the future absorption and exploitation of the scientific knowledge. This can produce an incomplete transfer of knowledge, which can instigate unnecessary rework and delay for industrialists.

IV. ASK FRAMEWORK PRESENTATION

The AsK framework aims to help a knowledge engineer to acquire and structure useful scientific knowledge into a k-base made available for future processing.

Materials used in AsK are related to knowledge sources and tools. Knowledge sources include the available documentary (publications, reports, slides, work sheets...) and human expert resources (interviews). Knowledge tools include techniques and software used for knowledge collection, visualisation or modelling (diagrams, frames, trees, maps, tables, matrices).

The method used in AsK is based on the assumption that to acquire and to structure relevant scientific knowledge, all the 7 corpora of knowledge (*Initial data, Problem, Hypothesis, Test, Results, Interpretation, Conclusion)* have to be examined.

The AsK framework proposes a three-phase protocol: i) acquire and structure knowledge from available resources, ii) identify the lacking necessary knowledge and then acquire it and iii) verify the conditions for acquisition completed.

During the first phase, knowledge is collected and modelled from the initial sources, following the subsequent procedures:

- The research problems for each knowledge domain are identified. Each problem introduces an empty AsK-frame entitled with regard to the given problem. This frame is a skeleton designed to offer a structured view on the scientific knowledge, as described in Section III. Each frame encloses seven layers, which correspond to the seven corpora of knowledge, as shown in Figure 1.
- Knowledge is collected from the initial sources and transcribed into suitable models. Glossaries, process diagrams, dependence relations, concept trees, process maps, attributes' tables, relation matrices, dependence rules are potential knowledge models (k-models) [8] to be put within the k-base. As far as possible, process-oriented models are used, as they ease the absorption by the industrialists. Knowledge modelling provides punctual views of the collected knowledge and delivers localised pictures within the k-base. The k-models reveal knowledge about concepts, hypothesis, methods, products, properties or results. They are placed into the corresponding AsK-frame, within the suitable layer. Knowledge collected and modelled is thus immediately located in its proper context. The procedure leads to a scientific-oriented structuring of the k-base.
- First interviews with experts are set up to confirm the initial AsK-frames titles with respect to their research domain. They can guide to the creation of new frames. This procedure provides the framing of knowledge (allows to assign one ore more experts to each frame, depending on the objective of their research work).

 ISBN: 978-1-61208-181-6

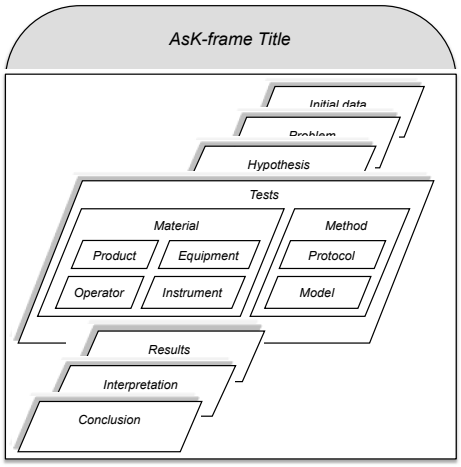


Figure 1. Seven necessary layers in the AsK-frame skeleton

- The initial established k-models are either confirmed or clarified, detailed or infirmed (if knowledge have evolved since then). New knowledge related to these models is collected and submitted to confirmation. The confirmed k-models are checked as relevant in the base
- The proper assignation of the confirmed k-models to a particular layer in the AsK-frame is confirmed as well.

During the second phase, the lacking useful knowledge is first identified and then acquired. The main steps to follow are:

- For a given frame, lacks of useful data, information and knowledge in the corpora of *Initial data, Product, Method, Equipment, Results* and *Interpretation* are identified with respect to the AsK framework.
- New semi-structured interviews are set up in order to collect these corpora of lacking knowledge within each frame. For example, the questions are formulated so that the knowledge obtained from answers directly fit a corpora (AsK-layer), which was lacking in knowledge.
- This new collected knowledge is modelled and the k-models are submitted for confirmation and checked following a similar procedure as cited formerly.
- The triple procedure (collect, model and confirm) is iterated for all the corpora of lacking useful knowledge requested for a given layer of an Ask-frame. Then, the procedure is iterated for all the layers within the frame and for all the frames created for the research domain.

During the third phase, the knowledge engineer verifies all the conditions for "acquisition completed", as drawn here:

- When all the corpora requested for a layer of an AsK-frame are collected, modelled and confirmed, the layer is considered "completed" and checked in the k-base.
- When all the layers of an AsK-frame are "completed", the AsK-frame is declared "completed" and checked.
- When all the AsK-frames within a given domain are "completed", the acquisition of the knowledge related to that domain is considered "completed".
- When all the domain acquisitions are "completed", the global acquisition procedure is declared "completed".

When the final condition is satisfied, the k-base is finished. It contains relevant and useful scientific knowledge, structured in order to ease the access for representation and reformulation.

V. AsK APPLICATION TO A BISCUIT-MAKING DOMAIN

We have applied the AsK framework to acquire knowledge in the domain of design and manufacturing of biscuits with satiety benefit. The work is in progress and it is carried out within a research project with industrial and academic partners.

The materials used as initial sources for collecting knowledge are available documentation (reports, presentations of domain research contributions, notes, and also transcriptions of recorded individual or collective interviews with experts). Three teams of scientific experts (in Formulation, Nutritional characterization and Sensory analysis) and an industrial partner (in the biscuit-making domain) take part in the project.

The method employed to acquire and structure scientific knowledge for each of the expertise domain suits the steps of the protocol mentioned in Section IV. Various k-models are built from initial sources and then submitted to the experts, for individual and group confirmation. The experts are interviewed at different stages during the acquisition process. Figure 2 illustrates one k-model built for dough texture's descriptors.

This k-model is confirmed and placed in the *Product* sub-layer of the AsK-frame "*Dough processability",* created for the Formulation domain. Additional and more specific knowledge is collected regarding to each of the descriptors of the dough texture shown in this k-model. This knowledge is modelled in well-formatted electronic files, which contain pictographic, textual and structure-related information (keywords, links and bibliography used). An example is given for the 'adhesiveness" descriptor, as pointed out in Figure 3. These electronic files are called knowledge sheets (k-sheets) and match the [8] meta-k-models. They are created using a specific computational tool that we have developed to assist the e-k-book implementation. Easy to be accessed, the k-sheets facilitate the knowledge visualization during the acquisition process. They will be properly treated and used as elements in the e-k-book.

K-models and k-sheets are established and then confirmed for all of the texture descriptors. Lack of useful knowledge for "*Dough Processability"* frame is identified with respect to the AsK framework. Then, corpora of knowledge for Hypothesis, Method, Instrument, Equipment and Results are acquired by means of k-sheet models to fill the layers of the given frame.

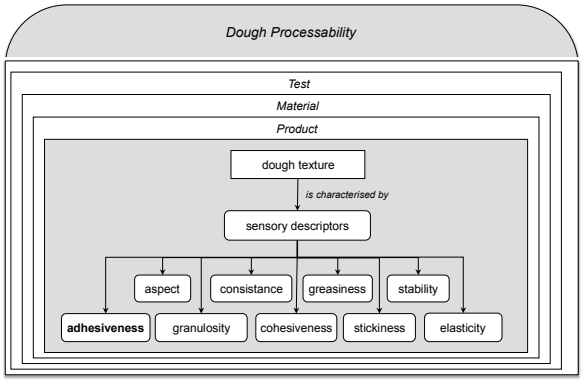


Figure 2. AsK placement for dough texture's characteristics map (k-model)

 ISBN: 978-1-61208-181-6

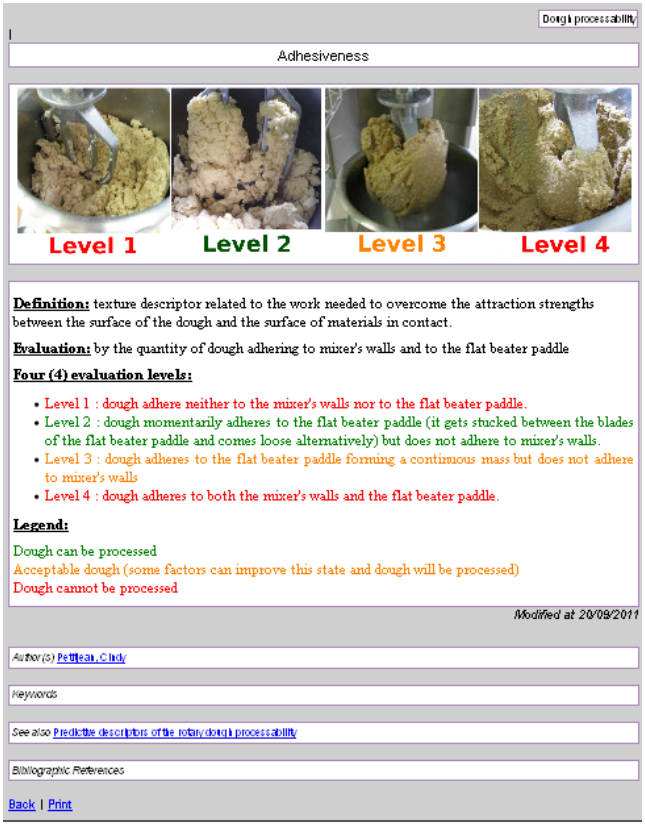


Figure 3. K-sheet collected for the "adhesiveness" descriptor

The example related to the evaluation of dough texture's sensitive descriptors is inspired by [11] and concerns published scientific findings. So far, two cases have been encountered in our project: i) scientific knowledge is already validated by the specific-oriented community and ii) scientific knowledge is in process to be produced and/or published (tests are in process, results are not yet available or published). In the first case, the scientific discourse is clear and it eases the acquisition process. In the second, the discourse is in construction and as a result, the knowledge acquisition can be significantly penalized.

VI. CONCLUSION AND FUTURE WORK

We have proposed here a framework (named AsK), which defines the seven corpora of knowledge (*Initial data, Problem, Hypothesis, Tests, Results, Interpretation* and *Conclusion*) that are necessary to be collected and modelled in order to acquire relevant scientific knowledge in the context of a U2I-KT [12]. With regard to these corpora, Ask offers a scientific-oriented [10] way to structure knowledge within the k-base, throughout the collection and the modelling process. Thus, when the acquisition procedure is completed, an organized k-base [8] containing all the necessary [9] corpora of useful and usable scientific knowledge [1] is made available for future treatments (industry-oriented reformulation, representation, and transfer).

The major advantages of using the Ask framework can be summarized as follows: i) it fits for multi-domain knowledge acquisition in parallel; ii) it offers to the knowledge engineer a strategy for planning and preparing the interviews during the acquisition procedure; it helps him to set questions during these interviews, which provide relevant knowledge, structured in a scientific-oriented view, as presented in scientific publications; iii) AsK supplies a global view over the k-base configuration and a granular view on a given k-model within the k-base; iv) it provides a helpful technique to instantly situate an already built k-model into its context within the k-base; v) Ask provides information about the vacant parts of the k-base; it detects lack of or not yet available useful knowledge necessary to acquire; vi) it gives information about the k-base local or global status (in process/completed) at any time of the acquisition process. Moreover, the Ask framework allows listing extra information potentially useful for further research (such as complementary methods to use, new hypothesis emerged, results to correlate). New research programs can be proposed and consequently new related knowledge can be generated, acquired and integrated in the (evolving-designed) e-k-book.

Further work will be dedicated to finish the acquisition process so that all the scientific findings will be available into a relevant k-base. Reformulation strategies and representation formats will be set up to make the acquired knowledge usable by industrialists and an e-k-book will be built for its transfer.

ACKNOWLEDGMENT

This work was carried out with financial support from the ANR France. The authors thank the BISENS project partners for all their contributions to the knowledge acquisition process.

REFERENCES

[1] T.H. Davenport and L. Prusak, Working knowledge-How Organization Manage What They Know, 1998, MA: Harvard BS Press, pp. 100-104.

[2] J.M. Beyer, "Research utilization: Bridging the gap between communities" in Journal of Management Inquiry, vol. 6, 1997, pp. 17–22.

[3] J. Kang, M. Rhee, and K.H. Kang, "Revisiting knowledge transfer: Effects on knowledge characteristics on organizational effort for knowledge transfer", in Experts Systems with Applications, vol. 37, n° 12, 2010, pp. 8155-8160.

[4] A. Agrawal, "University-to-industry knowledge transfer: literature review and unanswered questions", in International Journal of Management Reviews, vol. 3, issue 4, 2001, pp. 285-302.

[5] A. Hughes, "University-Industry Links and U.K. Science and Innovation Policy" in How Universities Promote Economic Growth (eds. S.Yusuf and K. Nabeshima), ch. 4, 2007, pp. 71-91.

[6] W.M. Cohen and D.A. Levinthal, "Absorptive Capacity: A new perspective on learning and innovation", Administrative Science Quarterly, vol. 35, 1990, pp. 128-152.

[7] J.H. Boose, "A survey of knowledge acquisition techniques and tools" in Knowledge Acquisition, vol. 1, n°1, 1989, pp. 39-58.

[8] N.R. Milton, "Knowledge acquisition in practice. A step-by-step guide", ed. Springer, 2007, pp. 69-110.

[9] C.B. Stetler, "Updating the Stetler model of research utilization to facilitate evidence-based practice" in Nursing Outlook, n° 49, 2001, pp. 272–278.

[10] J.Y. Cariou, "La formation de l'esprit scientifique: trois axes théoriques, un outil pratique: DiPHTeRIC", Bull. APBG, n° 2, 2002, pp. 279-320.

[11] C. Petitjean, P. Roussel, C. Simonnot, S. Berland, P. Aymard, and C. Michon, "Recherche des descripteurs prédictifs de la machinabilité des pâtes moulées", Poster Presentation, 60èmes JTIC, 2009.

[12] E. Rasmussen, O. Moen, and M. Gulbrandsen, "Initiatives to promote commercialization of University Knowledge", in Technovation, n° 26, 2006, pp. 518-533.

 ISBN: 978-1-61208-181-6

Web Services Integration with Regard to the Metrics of Data Believability

Adam L. Kaczmarek
Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology
ul. G. Narutowicza 11/12, 80-233 Gdansk, Poland
e-mail: adam.l.kaczmarek@eti.pg.gda.pl

***Abstract*—The paper is concerned with estimating the believability of data acquired from web services. In the paper, a new method for believability estimation is introduced. The method is designed for integrating web services. The believability estimation is based on the following metrics: quantity, reputation, approval, independence, traceability, maturity, authority and objectivity. In the method, data trustworthiness is determined by the credibility of the data source. In the believability estimation, information about data provenance is used. Moreover, the method is based on the consideration that it is possible to increase the data believability not only by finding more believable sources of data, but also by acquiring the same kind of data from many different sources and analyzing this data. It is possible in the field of web services because there are many vendors of the same kind of services.**

Keywords-data believability; web services; data processing

I. INTRODUCTION

When new data is acquired, the problem with data credibility occurs. Human beings practice various ways to estimate the trustworthiness of information and sources of information. There is a growing need to develop techniques, which can enable computer applications to automatically estimate the believability of information. These techniques should make it possible to identify and exclude unlikely information. It is particularly important in the case of applications collecting data from external sources through the Internet. These kinds of applications embrace software based on web services architecture [1].

This paper is concerned with data believability in web services. Applications using web services in order to achieve their goals receive data from services provided by vendors, which may not be reliable. It is particularly important for applications, in which integration of different kinds of web services is performed. If data collected by an application is wrong, the application will not produce proper results. Thus, it is necessary to verify the believability of data.

Data believability was defined by Wang and Strong as "the extent to which data are accepted or regarded as true, real and credible" [2]. The notion of believability is also referred to by the terms credibility, trustworthiness and plausibility. It needs to be stressed that data believability differs from data security. Security is related to problems of authentication, authorization and access to data. Data corruption may be caused by an undesirable influence of people or malicious programs as a result of poor data security. However, even if security is ensured and data is safely delivered from the source it is supposed to come from, problems with the truthfulness of this data may still exist: the source may spread incorrect information. Data believability is thus an issue, which goes beyond data security and it occurs even if problems with security are resolved.

The paper consists of five sections. The section following the introduction contains an overview of related work concerning data quality, believability and provenance. The third section presents the method for data believability estimation designed by the author of this paper. The fourth one contains an evaluation of the method. The last section refers to conclusion and future work.

II. DATA BELIEVABILITY

Estimation of data believability requires regarding of data not only by meaning, but also in the context of its provenance. Data provenance is an integral feature of data. On the basis of who created data, how it was stored, and how it was processed, conclusions can be drawn about data believability. Recently, there has been a significant amount of research on acquiring and storing information about data provenance in web services. Tsai et al. presented a profound description of requirements and solutions concerning data provenance problems [3]. Techniques for solving these problems include the use of metadata, databases and new types of protocols. Moreover, an in-depth description of data provenance problems was presented by Moreau [4]. Storing information about data provenance makes possible to use this information in order to estimate data believability.

Information about data provenance indicates a web service, which is the source of that data. The quality of this web service can be taken into account in estimating data believability. Some web services' parameters are objective and they can be determined on the basis of statistics about the behavior of web services. Such parameters include web service availability, fees and latency in data transmission [5]. The quality of a web service can be also evaluated by its users in the same way as on eBay or Amazon, where customers provide their feedback about products and suppliers. It is possible to prepare ratings concerning the quality of web services. Such ratings can be based on both objective metrics, referring to the performance of a web

 ISBN: 978-1-61208-181-6

service and the opinions of the service users. In the field of data believability estimation, the most important are ratings concerning users' feedback about trustworthiness and believability of data delivered by web services. Various methods for rating web service quality and managing trust in web services were described by Golbeck [6].

Another important problem concerning data believability estimation is determining the metrics of data quality. Wang and Strong wrote an influential paper, in which they presented a detailed list of data quality attributes [2]. They also specified 20 dimensions characterizing data quality. Problems with measuring data quality are also the main subject of a book by Khan [7]. Furthermore, a paper [8] wrote by the same author as this paper, presents a method for evaluating the credibility of information in semantic web and knowledge grid.

III. A NEW METHOD FOR RATING DATA BELIEVABILITY

This section presents a new method for rating data believability. It was designed by the author of this paper and it is intended to be used in the integration of web services. The method consists of determining the level of data believability. This level is calculated on the basis of multiple metrics.

Common methods for rating data believability focus on determining the believability of individual sources, in order to select the most believable information provided by one or other of those available sources. The method introduced in this paper represents a different approach. It is designed to acquire the same kind of data from many different sources. On this basis conclusions about data believability are made. In the method, a level of data believability is calculated. The value of this level can be higher than 1. The level of believability is not like the probability of data trustworthiness. It is a perceived believability estimated by the method. In the case of one source of information, the value of the level ranges from 0 to 1. When there are more sources the level can be higher than 1.

In order to define metrics, which indicate the level of believability, a distinction between a claim and data, needs to be made. When some source of information publishes data, it cannot be assumed that this data is definitely true. Data provided by sources will be, in this paper, called claims, similarly as in [9]. Claims are also a kind of data, but there are two additional features of a claim:

- The source of a claim is specified.
- It is not resolved whether a claim is true or false.

Claims will be, in this paper, denoted by the symbol ζ. The data corresponding to claim ζ, but without a specified source, will be denoted by d_ζ.

It also needs to be considered what the granularity of data, being a claim, is. In the method, it is assumed that a claim is a portion of data of any size. A claim can be either one logical sentence, a sequence of such sentences, or the whole portion of data received from a web service. The size of the data does not affect the process of determining its believability. The only difference is that the level of believability concerns different data. The general formula for calculating the data believability level is given by (1).

$$b(d_\zeta) = \sum_{i=1}^{m_0} \left(\left(\frac{\sum_{k=1}^{M} w_k m_k(\zeta_i)}{M} \right) |\zeta_i| \right) \quad (1)$$

where b is the believability level, m_0 is the number of sources supporting claim ζ, which corresponds to data d_ζ, symbol ζ_i refers to claim ζ announced by source i, symbol m_k denoted metrics used in calculating the believability level, w_k is the weight of a metric m_k, letter M stands for the number of considered metrics, letters i and k are indexes used in additions. Symbol $|\zeta_i|$ in (1) represents the influence of a single claim given by the source number i to the level of data believability without the use of any metrics. Expression $|\zeta_i|$ is similar to the cardinality of a set. Every claim stands for a single portion of data, so the value of $|\zeta_i|$ is always equal to 1. This kind of notation is used in order to indicate that only claims, which correspond to data d_ζ, have influence on the level of this data believability.

Equation (1) states that the level of believability is equal to the number of claims supporting data d_ζ with regard to the metrics. Each source providing claim ζ_i increases the believability level of data d_ζ. The extent of this increase is equal to the weighted average of all metrics concerning the claim. Weights used in calculating the weighted average correspond to the importance of the metrics.

The following metrics are taken into account: quantity, reputation, approval, independence, traceability, maturity, authority and objectivity. The values of all metrics used in the method presented in this paper, apart from metric quantity, range from 0 to 1.

In the method presented in this paper, data believability is estimated on the base of attributes of the data source. There is also a possibility to estimate data believability on the basis of the data content. Several attributes of the data itself can be taken into account; like data validity, its accuracy and the context to which the data applies. In the method presented in this paper it is not considered. Data is only evaluated on the basis of the believability of its source. Nevertheless, it is possible to enhance the method with these attributes.

A. Quantity

The quantity of claims acknowledging the data being under verification is denoted with m_0. Although this value is not present in (1) in the same way as the other values of the metrics, it in fact refers to one of the metrics considered in determining the believability level. The value of m_0 defines the upper bound of the first summation presented in (1). The metric quantity represents the principle that the more sources are announcing the data, the greater is the level of believability. An assumption was made that the relationship between the number of claims and the level of believability, is linear when no other metrics are taken into account. All sources are then treated equally. The believability based only on the metric quantity is presented by (2).

 ISBN: 978-1-61208-181-6

$$b_{m_0}\left(d_\zeta\right)=\sum_{i=1}^{m_0}\left(\left|\zeta_i\right|\right)=m_0 \quad (2)$$

where b_{m0} stands for the level of believability when quantity is the only metric considered. In this case, the believability level is equal to the number of sources providing claims corresponding to data d_ζ.

B. Reputation

In this method for determining the believability level, sources are not treated as if they were equal. Their influence on the level of believability is biased by various factors. First one is the reputation of the source. The value of the metric's reputation is denoted by m_1.

The reputation of a web service is defined as "a general opinion i.e., it aggregates the ratings of the given service by other principals. Typically, a reputation would be built from a history of ratings by various parties" [10]. Principals are understood here as service providers or requesters. In the method for estimating believability presented in this paper these kinds of opinions are taken into account in the form of a metric *reputation*. The value of this metric is calculated on the basis of a web services' rating system. There are a large variety of such systems. However, they are most often concerned with many web services' parameters, such as performance, reliability, latency, fees etc. The reputation concerned in this paper is based only on the opinion about the believability of data provided by a web service.

The level of reputation can be acquired from web services collecting data about the quality of other web services and web services' ratings. The quality of web service, in the context of data believability, needs to be given in the form of a numerical rating (e.g., 6 using a scale 0 to 10). Such a rating is converted to scale from 0 to 1 and it is directly used as a value for a metric reputation. In the presented method the metric reputation can also be based on ratings acquired from many sources and rating systems. In this case the value of the metric is equal to the average of ratings adjusted to a scale from 0 to 1.

C. Approval

Approval is a metric, which is similar to the metric *reputation* in the way that it is also concerned with the behavior of a web service in the past. Whereas metric reputation is indicated by third parties, metric approval is the own opinion of a customer of a web service. The customer can, and should, store data about cooperation with web services. In the case where a web service provided data, which appeared to be wrong, the believability of this web service is decreased. On the other hand, the believability of proven web services should be increased.

The metric approval has three parameters: *q, p* and *T*. Parameter *p* is the influence of providing by a web service appropriate data, parameter *q* represents the impact of publishing wrong data and parameter *T* is concerned with the time, after which the influences caused by wrong and right data are no longer valid. It is assumed that the change in the value of metric approval is not perpetual and after some time the impact of providing wrong, or right, data is eliminated. The impact is diminished linearly, starting from the initial level of parameters *p* and *q* until there is no influence when time T has elapsed. The formula for calculating the metric approval is presented by (3).

$$m_2=m_{2base}\prod_{k=1}^{N_q}\left(q+\frac{t_k}{T}(1-q)\right)\prod_{l=1}^{N_p}\left(p+\frac{t_l}{T}(1-p)\right) \quad (3)$$

where m_{2base} is the default value of metric approval, N_q is the number of valid influences of providing wrong data, N_p is the number of valid influences of providing right data, q is the initial level of influence for wrong data, p is the initial level of influence for right data, t_k and t_l are times since the event of providing wrong or right data occurred and T is the parameter indicating the time, after which occurrence of wrong or right data is no longer taken into account.

The values of m_{2base} *p, q* and *T* included in (3) need to be specified. Parameter m_{2base} is the value of the metric when there is no experience in cooperation with the web service. The attitude to such a web service is neutral. Thus, the value of that metric is then equal to 0.5. In setting the value of parameter q it needs to be stated how severely the believability of a web service should be diminished when a web service provided wrong data. The parameter q can have various values. For example, it can be assumed that when no other metrics are taken into account, two web services, which once provided wrong data are as believable as one unknown web service. In this case, the value of metric approval is reduced by a half when a web service provides wrong data. When data acquired from a web service is right, the value of metric approval can be increased by half of its previous value. Thus, possible values of parameters are q=0,5 and p=1,5. For example, a web service, which once provided wrong data and once right, would have the level of metric approval equal to $m_{2base}\times p\times q$=0,5×0,5×1,5=0,375.

The value of parameter T can be selected arbitrarily and it can be set to 365 days. Values of t_k and t_l can be then changed once a day. They would indicate the number of days since wrong, or right, data was extracted from a web service.

D. Independence

The metric *independence* arises from the remark that data confirmed by two independent sources is more believable than data provided by two sources when one of those sources obtains data from the other one. A similar rule is applied by press agencies, assuming that information is true when it is confirmed by two independent sources.

In the method presented in this paper, the metric independence indicates the number of independent sources of data. If there is only one source of some information and other web service providers supply data on the basis of that one source, the value of metric independence is as minimal as possible. It is then equal to 0. On the other hand, a maximum value of metric, i.e., 1, indicates that there is an unlimited number of independent sources. In order to satisfy these conditions, the metric's value is increased because of subsequent independent sources in a similar way as a

 ISBN: 978-1-61208-181-6

geometric progression. The value of metric independence is presented in (4).

$$m_3 = 1 - \frac{1}{a^{u-1}} \quad (4)$$

where m_3 stands for the metric independence, a is the parameter determining the level of the increase caused by the existence of subsequent independent sources and u is the number of independent sources providing data. The value of a can be set to 2. Then, the medium value of the metric would mean duplication of data by two independent sources.

The metric independence also applies to situations more complicated than repeating data provided by an independent source. Web services are based on acquiring data from various sources. When there is a group of web services, there is also a group of independent sources, from which data is acquired. In the group of web services some part of data may be derived from a smaller group of independent sources than the other parts. Different parts of data can be confirmed by a different number of independent sources. There is a part confirmed by the smallest number of sources: the number of a source confirming this part is assigned as the value of parameter u from (4). Thus, when there is some data, which all web services acquired from the same source, the value of parameter u is set to the same level as if there was only one independent source. The value of u would be equal to 1.

E. Traceability

Metric *traceability* is another metric used in the method presented in this paper. The value of the metric is denoted by m_4. It depends on whether a web service specifies sources of information, which were used to make the service available, or there is no such information. If web services base their results on data acquired from other parties, they should provide information about that. For web services preparing all data by themselves, there should also be a notice that no external sources were used. Providing data about sources of information is possible due to the researches concerning storing information about data provenance.

Providing source information is similar to the bibliographies presented at the end of scientific manuscripts. When bibliography is not present or it is poor, a paper is treated as less credible. When a web service does not provide any data about its sources of information, the value of metric traceability is the smallest possible, i.e., equal to 0. If full information is available, the value of metric traceability is equal to 1. It is also possible that information about sources is partly present. In this case, metric traceability corresponds to the extent of source information availability.

F. Maturity

Another metric used in the method presented in this paper is *maturity*. The metric is based on the premise that web services, which are operational for some period of time are more believable than those, which are new and not tested by customers. Similarly, companies with tradition are more respected than the new, and unproven, ones.

The value of the metric maturity depends on the time elapsed since the release of a web service. The time, after which a web service is treated as fully believable, can be determined differently. It can be assumed that the believability of web services, which are available for over one year, is no longer reduced by the metric maturity. Web services, which are absolutely new, have the lowest value of metric maturity equal to 0. The values of this metric are changed linearly, for those web services whose time of service is in between these limits. Thus, the value of metric maturity is given by (5).

$$m_5 = \min\left(\frac{t}{T}, 1\right) \quad (5)$$

where t is the time, which elapsed since the release of a web service and T is the period of time, after which the believability of a web service is not reduced by the metric maturity. The parameter T can be set to 365 days and the value of the parameter t can be changed daily.

G. Authority

The value of metric *authority* is denoted by m_6. This metric refers to the sources' competence to provide data. In particular, it concerns data that sources claim to have obtained themselves. In the case where there is a doubt that the source does not have qualifications to provide a certain kind of data, the value of metric authority is reduced. For example, when a web service provides data concerning the number of people on Earth, claiming that this data was acquired by itself, there is a reasonable basis to distrust such information.

It is problematic to estimate the value of metric authority. Methods of storing provenance data do not reach a complex enough level to correlate the possibility of sources of information with data they provide. The value of metric authority needs to be assessed partly manually. In particular, this metric would concern data that only some kinds of sources are able to obtain. For example, information about population in countries can mainly be derived only from government sources. The value of metric would be equal to 0 for sources, which do not satisfy the requirements and it would be equal to 1 otherwise. A list containing specific kinds of data with corresponding sources can be prepared. When a list is available, applications using the believability estimation method presented in this paper can automatically use this previously prepared list.

H. Objectivity

The value of metric *objectivity*, denoted with m_7, is in most cases equal to 1. This value is changed for data that can be biased by the source due to its own interests. The value of metric objectivity is lowered for information that vendors claim about their products and their quality. Companies, on the basis of marketing needs, tend to modify information in order to improve their image. In such cases the metric objectivity is set to 0, as the source is not objective. It is possible to determine the value of this metric automatically

 ISBN: 978-1-61208-181-6

on the basis of metadata concerning product, their manufactures and resellers. In case no such data is available, these metrics can be set on the basis of a manually prepared list similarly as with metric authority.

I. *Weights of metrics*

The influence of metrics on the level of believability is modified by the weights of these metrics. Some metrics are treated as more important than others. Apart from metric quantity there are seven metrics with weights. The sum of weights has to be equal to one, because when there is only one source of information and all metrics are equal to one, then the level of believability needs to be also equal to one.

The most important metrics are reputation and approval. In fact, conclusions about the web service believability can be drawn only on the basis of own opinion about a web service and opinion of others. Moreover the values of other metrics in many cases will be the same for different web services. The weight of metrics reputation and approval need to be higher than other weights. Weights of metrics reputation and approval can be set to 0,25 and weights of other metrics can be equal to 0,1. Thus, $w_1=w_2=0.25$ and $w_1=w_2= w_3=w_4= w_5=0.1$.

J. *Application of the method*

There are three types of information sources used when the method is applied: web services, third parties and own knowledge. Third parties provide information about the quality of web services, such as their reputations. When an application needs a certain kind of information it collects claims from different web services concerning this information. It is like stating a question and seeking for the answer. The application also collects information from third parties and takes into account its own knowledge. The level of believability of each kind of the answer is rated with the use of the method. As the result, the answer with the highest level of believability is regarded as truthful. Acquiring data from many sources is more expensive then taking into account only one source, however the idea of the method is to improve the quality of data despite increased cost.

IV. EVALUATION

The method is based on the assumption that in general information provided by web services are truthful. The method resolves the problem of excluding information from untruthful, low quality web services (on the basis of metrics reputation, approval, maturity, authority and objectivity). It also manage the problem of providing wrong data by a noble web service due to some accidental mistake (on the basis of metrics quantity). In such cases, without using the method, false information would be regarded as truthful.

However the method does not guarantee the truthfulness of information. In case false information is universally regarded as truthful the method will also accept the truthfulness of information. Nevertheless the method attempts to disregard such information (on the basis of metrics independence and traceability). In case of such information there are some sources, which published it. If information concerning data provenance were commonly provided the spread of untruthful information could be limited. The results of the method in case of this kind of untruthfulness depend on the availability of information about data provenance.

V. CONCLUSION AND FUTURE WORK

The method presented in this paper makes it possible to automatically estimate the data believability level on the basis of information about data provenance and web services' ratings. In further work, we are planning to enhance the method with metrics referring not only to the believability of the source of data but also to the data itself.

One of the significant problems related to the presented method is that web services should provide information about sources of data, which were used to make the service available. This would protect other applications from propagation of wrong data in case some source is publishing not truthful information.

ACKNOWLEDGMENT

This work was supported in part by the Polish Ministry of Science and Higher Education under research project N N519 172337.

REFERENCES

[1] D. Booth, H. Haas, F. McCabe, E. Newcomer, M. Champion, C. Ferris, and D. Orchard (eds.), Web Services Architecture, W3C Working Group Note 11 February 2004, W3C, 2004.

[2] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers" Journal of Management Information Systems Vol. 12, No. 4, , Spring 1996, M. E. Sharpe, Inc., pp. 5-34.

[3] T. Tsai, X. Wei, Y. Chen, R. Paul, J.-Y. Chung, and D. Zhang, "Data provenance in SOA: security, reliability and integrity," Service Oriented Computing and Applications, vol. 1, no. 4, Springer-Verlag, Dec. 2007, pp. 223-247.

[4] W. Moreau "The Foundations for Provenance on the Web" Foundations and Trends in Web Science, Now, 2009, submitted for publication.

[5] M. Ouzzani and A. Bouguettaya, "Efficient Access to Web Services" IEEE Internet Computing, Mar./Apr. 2004, IEEE Computer Society, pp. 34-44.

[6] J. Golbeck, "Trust on the World Wide Web: A Survey," Foundations and Trends in Web Science, Vol. 1, No. 2, Now, 2006, pp. 131-197.

[7] K. M. Khan (ed.), Managing Web Service Quality: Measuring Outcomes and Effectiveness, IGI Global, 2009.

[8] A. L. Kaczmarek, "Automatic Evaluation of Information Credibility in Semantic Web and Knowledge Grid, "Proc. of the 4th International Conf. on Web Information Systems and Technologies (WEBIST 2008), vol. 2, Funchal, Madeira-Portugal, INSTICC, May 2008, pp. 275-278.

[9] C.Bizer and R.Oldakowski, "Using Context- and Content-Based Trust Policies on the Semantic Web," WWW 2004, New York: ACM, May 2004, pp. 228-229.

[10] E. M. Maximilien and M. P. Singh, "Conceptual Model of Web Service Reputation" ACM SIGMOD Record: Special section on semantic web and data management, Vol. 31 , No. 4, Dec. 2002, pp. 36-41.

 ISBN: 978-1-61208-181-6

A Comprehensive Study on the Reality of Knowledge Management and Lessons Learned in the Projects

A Case Study in Iran Oil and Gas projects

Ahad Nazari
Assistant Professor, Faculty of Architecture and Urban Planning, Shahid Beheshti University
Tehran, Iran
Nazari_ahad@yahoo.com

Mohammad Mehdi Mortaheb
Assistant Professor, Civil Engineering Department, Sharif University of Technology
Tehran, Iran
mmmortaheb@yahoo.com

Zahra Aghalou
Engineering Department, Ministry of Petroleum, Expert on Project Management
Tehran, Iran
aghaluz@yahoo.com

Abstract—**The concept of Lessons Learned (LL) in projects refers to the knowledge and experiences have been gained during the execution of the projects. These can be very vital in improvement of management style of current and future projects, to prevent the potential problems such as cost and time overruns. However, despite the importance of LL, it is usually ignored or not considered properly in the projects. The aim of this research is to study the current status of the LL practices in "Iran Oil and Gas Projects" (IOGP). First, a comprehensive literature review on the concept and approaches used to include LL in project management is conducted. Then, the major barriers for implementation of LL management in IOGP are identified, using interview method and investigation of the current states of LL processes in IOGP. Finally, in order to suggest solutions, in-depth analysis of the problems is carried out, using project management tools and techniques. According to the findings of this research, there is not a formal and systematic process for LL management in IOGP. Moreover, the main barriers for KM and LL Process consist of strategic and management barriers, organizational barriers, communication barriers and staff barriers. The proposed solution to improve the LL documentation consists of organizational issues, open communication culture, training & learning environment and formal LL process.**

Keywords-Project management; Lessons learned; Knowledge Management; Oil and Gas Projects

I. INTRODUCTION

Completed projects are terrific sources of information [3], which bring some new experiences on the table. Such information can help to prevent similar problems in future and make considerable time and cost savings on the projects. The concept of Lessons learned (LL) in projects refers to the knowledge gained from the process of performing the project. It can be documented at any time of project life cycle, particularly during the execution of the projects, in order to catalogue significant information that has evolved as a result of the implementation of the projects. The gathered information is used to build up a knowledge base of an organization and to establish a database of the best and worst experiences in project implementation [1]. LL has a significant role in knowledge development and improvement, particularly in uncertain problems. It provides the possibility of recording the knowledge gained by professional bodies who have worked thousands of hours in the projects. So even if they leave the organization, it can be documented and reported as the LL knowledge base, which is a database of historical information, gained experiences and issues related to the outcomes of previous projects [2].

Despite the importance of KM & LL, they are usually ignored or not considered properly. This deficiency is common not only in many of the projects in the world but also in the "Iran Oil and Gas Projects" (IOGP). So, the aim of this paper is to study the current status of LL documentation in IOGP with concentration on its barriers. According to this aim, through a comprehensive literature survey on the concept of LL and KM, the processes, tools and requirements for developing a successful LL process is introduced. Then the barriers for KM and LL process are identified. Finally, through some interviews and documents analysis, the status and the barriers for implementation of LL process in IOGP is investigated and some solutions are proposed.

II. LITERATURE REVIEW

A. KM and LL Reality in the Projects

Lessons learned are the experiences gained from the process of performing the projects. It is an ongoing process with great knowledge creation. It may be considered as a project record, which is classified as a lessons learned knowledge data base [2]. The project team may learn lessons which can be useful in similar future projects, while no one documents them in a systematic way at the end of the projects [11]. Using knowledge gained from implementation, failures or even successes of the projects is vital for the long-term sustainability and competitiveness of businesses [5]. It can be helpful in; reduction of the cost and time needed for problem solving and improvement of the quality of the solutions suggesetd

 ISBN: 978-1-61208-181-6

during the construction phase of the projects. Morover, it can reduce the probability of repetition of the same problems in projects [12]. So, most of the projects do not need to start from scratching as much as they can utilize existing processes and learn from the experiences acquired from previous projects [5].

According to "integration management" part of PMBOK [2], project close out refers to the process related to the closing of the projects including LL documentation. This document includes *"Historical Information"* which is transferred to the *LL knowledge* base for use by future projects. It can be concluded that LL activities is a major part of the knowledge management in the projects. This information is used to update "Organizational Process Assets".

Knowledge Management (KM) is the process of creating value from an organization's intangible assets. It refers to sharing and leveraging knowledge within an organization and outward toward customers and stakeholders [5]. KM has such strategic value that organizations should include it as one of the pillars of their human capital strategy. KM can help to capture, share and leverage knowledge before it leaves the organization [5]. KM improves the efficiency of the organizations; it facilitates the access to the knowledge of the employees. The organizations make better decisions, improve their procedures, reduce reworks, increase innovation and reach to high level of integration and cooperation [6].

B. KM and LL Processes and Tools

Several KM cycles are suggested in different references, while the simplest one that meets the purpose of this paper consists of four major steps, Knowledge is identified and captured, shared with others, applied in combination with existing pertinent knowledge, and then created in the form of new knowledge, which is then captured and continues [5].

According to Chin, (2004) [13] There are three parts to the basic LL process. First, develop an environment that supports continual learning. Then, capture key lessons learned and finally, archive, organize, and make these learnings accessible to current and future project teams. So, archiving, organizing, and communicating LL will be a foundation for long-term success of the organizations.
Considering the aim of this approach, it is possible to create major three steps for LL process including; LL capturing approaches which refers to the activities of capturing key lessons learned, LL documenting which refers to the activities of archiving and organizing and LL communicating which refers to the activities of making LL accessible to current and future projects. Comparing the KM cycle as explained previously with the above suggested steps for the LL process, It can be concluded that the concept of the steps is similar but as it mentioned the fourth step of the KM cycle is use of knowledge by people and then it may result in creating the form of new knowledge which is then captured and continues, therefore for completion the cycle of the LL process another step should be added as "LL Feedbacks & Development". So it implies that LL process should include four major steps: Capturing Phase, Documenting Phase, communicating Phase, Feedbacks & Development Phase:

C. Capturing Phase

The purpose of "LL capturing" is to find out key experiences, deviations and any LL regarding the ongoing projects. Key information about achieved experiences including deviations, problems, opportunities, wrong or right actions/ solutions and root causes of any deviations and problems are captured from key projects' team members and Stakeholders. It should be noticed that there are some approaches for capturing of LL, experiences and root causes analysis, as below. It means that LL approaches should not be just a diary report of what happened in the project. Below, different approaches for "LL capturing" are presented [13].

1) Learning Organization Establishment: A learning organization is one in which managers do everything possible to maximize the ability of individuals and groups to think and behave creatively and thus maximize the potential for organizational learning to take place. In order to establish such environment, a learning organization should be built up. The five principles for creating a learning organization are to develop personal mastery, build complex, challenging mental models, promote team learning, build shared vision and encourage system thinking [10]. Organizational Learning (OL) is complementary to KM. An early view of OL was "Encoding inferences from history into routines that guide behaviour"(Levitt and March 1988). So OL has to do with embedding what has been learned into the fabric of the organization [9].

2) Meeting Approaches: There are different techniques to manage LL meetings such as brain storming [13], open discussion [14] and even E-meeting [4], in which all participants can discuss from their point of view in turn. The facilitators of the meetings should try to keep discussions on the road and un-biased. Well documenting and recording of the meetings' results is important [13]. The input data for these meetings include Project Schedule, Bug Reports, Review Reports, and Integrated Project Plan [15].

Generally LL meetings should be held bi-weekly, monthly or at major milestones with participation of the project team and stakeholders, as an information sharing tool of how obstacles were overcome and what could be done better on the next phase or next project. The advantage of regular meetings during the project progress (not just at project completion) is that people can give more accurate information and have better discussion, before forgetting the details of experiences or even leaving the project or organization.

- Brainstorming Technique: In this technique, the facilitator asks the participants to write their answers/ideas to questions such as: "what are the

 ISBN: 978-1-61208-181-6

main LL / experiences in the project", "what could be improved?" and "what went well?" In order to avoid people influence on each other point of views and cause biases on the results, it is recommended to develop the ideas in silence. Then the facilitator leads the discussion to record and publish the results of LL [13].

- Open Discussion: In this approach all participants can present issues including experiances, problems and their solutions. The LL meetings can be held in three levels, including project team members, relevant department managers and senior managers. So, team members feel free for open dialogue. The idea is to transfer the results of lower level meetings to the upper ones [16].

3) Capturing Tools

- IT Tools -Telecommunications & Media Tools**:** This approach to LL Capturing can be used especially when project members and stakeholders are not in the same place and can not have regular face to face LL meetings. Using IT facilities and media devices, such as e-mail, SharePoint, video conference and other tools, project managers are contacted in a specific time periods to find out if they have encountered similar problems and have knowledge of the solutions.
- Risk Management: According to the risk management part of PMBOK standard, it can have a significant role in LL capturing [2]. The results of risk management processes produce information that can be used in future projects and should be captured in the organizational process assets [2]. One of the organizational process assets that referred here which may be updated is lessons learned from project risk management activities. These documents are updated, during the implementation of the project and the project closure. Risk management plan, risk breakdown structure and the final version of the risk register, all include some issues related to LL documents [2].
  - Project Auditing: Project auditing can be addressed as a capturing tool for the projects' LL. Its report includes LL information. The formal report should contain Information pertinent to other projects, which means, what LL from the project being audited can be applied to other projects being undertaken by the organization [17].

D. Documenting Phase

As it is mentioned before, all efforts to capture LL data through each of the aforementioned methods would have no value if they are not documented in a well-organized knowledge base format. The following approaches are suggested for documenting the LL.

1*) LL Report:* LL data should be distinguished and extracted from resources such as LL meetings, risk management process and project auditing. The context of the LL report includes [1]: description of the LL, sources of LL data, any reference related to the LL, description of root causes of the problems, impacts of the problems on the project, recommended solutions, description of the application of the LL, searchable key words related to the LL documents, list of the necessary data for creation and reporting and approving the LL.

2) *Approving and Encoding the LL Documents:* Considering the main attributes of the LL report, the outcomes of LL reports need to be reviewed and categorized in a final report format, which is an appropriate reposition for the knowledge of informally developed procedures that helps to ease the tasks of execution of the project. Once reported, they can be tested, approved and finally added to the parent organization project management procedures and roles, if generally useful [17].

All generated LL documents should be approved by an entitled person of the organization that should be clarified in the LL procedure in advance. The Project Manager should assign some individuals at the early stages of the project for the task of generating these documents. Moreover, it is very important to enable the LL documents by using some searchable functions such as key words, their application areas, the level of their importance and chronology. So they should have label of encoding which shows these characteristics.

E. Communicating Phase

According to the philosophy of the LL documentation, information should be communicated with the relevant people in the organization in a timely manner and distributed to them. So, the communication process of the organization has a key role to ensure that LL is available to all relevant people, in a timely manner. According to the PMBOK standard [2], communication management process can facilitate LL process in term of gathering and distributing of LL from and to the relevant people. Both LL process and communication process increase the efficiency of each other and their functionality. There are many tools for communication improvement such as team working, communication media and so on.

There are some factors to be considered including Communication network, Communication climate and Trust Building [8]. Moreover, media is the way that information is communicated. Using a combination of some of the media such as Electronic database (information systems), Web Pages, Broadcasting Job Knowledge, email, Inter LL meetings, Journal and newsletters can help to communicate LL of the projects in an organization.

F. Feedbacks and Development Phase

Application of the LL is the main purpose of doing LL process in projects but it is not independent from other steps. LL application in other projects and similar situations, getting feedback and improving LL process are the most important part of a LL process. Feedbacks can be

 ISBN: 978-1-61208-181-6

reflected in order to update LL database. The mechanism of reflecting feedbacks can be considered as below:

- Risk Monitoring & Control: According to the Risk Management process, risk monitoring & control has a key role in updating the organization LL document. Risk management feedbacks are reflected to update LL data.
- Organization Procedures and Standard Adjustments: Normally, LL outputs could recommend some improvments on the technical and managerial assets of the organization. The recommended revisions will be investigated and will be added to the organization procedures and standards. This approach guarantees useful application of the LL findings which means LL data are communicated and applied in its relevant areas.
- Project Management Office (PMO) Establishment: Establishment of the Project Management Office (PMO) in organizations can be a guarantee to develop a successful LL process. PMO is an office to deal with multiple projects and charged with improving the project management maturity and expertise of the organization, as well as increasing the success rate of projects. PMOs commonly perform many tasks such as initiating and launching new projects, establishment and enforcing project management processes. Most of its functions cover different activities required to perform the four major steps of the LL process in an organization [17].

III. LL AND KM BARRIERS

According to the aim of this paper, in this section the general berries of KM and LL activities is reviewed. Considering the relevant literature, the main barriers of KM can be classified as below [7]. These barriers are valid in LL process and can affect its process as well.

A. Strategic and Management Barriers

- Strategy Alignment**:** The problem will occur due to the lack of alignment between the KM strategy and the business strategy of the organization.
- Management Support**:** Without the active and visible support from top managers in organization, KM will not get the support of employees. So, where leaders clearly communicate and enforce the value of sharing knowledge, the KM will have better results. One of the main management barriers is that managers do not use positional and social power to facilitate open communication channels in an informal way.
- Resource Allocation**:** The cost of capturing, processing and transferring knowledge can be a barrier to KM. When companies fail to allocate sufficient resource to the KM activities in most instances they will fail in their KM venture [18] [19].

B. Organizational Barriers

- Organizational Structure**:** Informal connections keep communication channels open and nobody needs to wait for next official meeting to initiate information. So, organizational hierarchy could have a direct impact on the KM success. In some organizations hierarchical structure negates any KM activities, which means organizations should be aware of that.
- Organizational Culture**:** The Company's culture can influence the perceived usefulness, importance and validity of KM. It influences the process of creation and adoption of new knowledge, determines the knowledge belongs to organization or individuals and creates a content of social interaction which influence**s** the organizational maturity on KM. Factors such as creating a supportive and open communication climate, fair rewarding and building trust in an organization can encourage people to state their beliefs and focus on opinions and problem solving rather than negative evaluations or criticizing others [8]. Where the organizational culture is not aligned with the drive for knowledge it will act as a barrier toward KM [7].

C. Communicational Barriers

- Communication and Information Systems**:** If the organization's information system is a kind of overloaded system (including useful and useless information) people will be confused [20]. People simply like to directly access to the required knowledge in an appropriate time scale with less confusion about the accuracy of the system. One of the major barriers to LL utilization is inefficient access to the relevant information. The communication process of the organization has a key role to ensure that LL is available to all relevant people.
- Rewarding and Recognition System**:** when the organization doesn't have a rewarding system to recognize the valuable knowledge shared by the employee, people do not encourage sharing valuable knowledge.

D. Staff Barriers

- Trust**:** Tacit knowledge is the type of knowledge that lies with the experience of employees, so in order to share this type of knowledge employees should trust the organization. Otherwise they will feel that they lose their own knowledge and its value, so they avoid making it available to others.
- Competition: If the competition internally in an organization or externally between divisions is not healthy, people will not want to share knowledge.
- Knowledge Value**:** The value and benefit of existing knowledge is often not realized by the owners, so

 ISBN: 978-1-61208-181-6

they don't try to capture and store it in an appropriate format.

- Language: Due to communication barriers, knowledge may not be transferred to the relevant people. This problem particularly can occur in multi-national projects which are implemented by many people from different countries [21].
- Staff Turnover**:** As discussed before, it is difficult to capture and manage tacit knowledge. This type of knowledge is also lost when employees leave the organization.

IV. LL CONCEPTUAL MODEL

Considering the above discussion about the LL elements and their relation with KM, a conceptual model is developed. The Model shows the barriers to KM and LL process in the organizations. Moreover, it presents the relationships between KM and LL process and their interactions with the projects (Figure 1). According to this model, the LL process is conducted during the implementation of the projects. The adopted data from LL are transferred to the knowledge base of the organization in an appropriate format, which can be used to support new projects. The model shows two main parts of LL and KM process in an organization and their relation with projects.

In the first stage, the LL is captured and documented that can be considered as captured knowledge in KM system. The next stage is LL communicating that is equivalent to the knowledge sharing stage of the KM process. Therefore any communication procedure**s** of LL process can be transferred for utilization in the knowledge sharing step as well. The last stage of LL process is application of LL in other projects and similar situations and their feedbacks which help to improve and update the LL database. This stage corresponds with knowledge application. Finally knowledge creation stage of KM process enhances knowledge capturing stage. During this process many problems and challenges could happen, which are classified as LL and KM barriers

V. INVESTIGATION OF THE LL BARRIERS IN IRAN OIL & GAS PROJECTS

In order to find the current status and the main barriers to documentation and utilization of the lessons learned, in Iran Oil & Gas projects, the actual application of LL in Iranian oil & gas projects is investigated. This case study has been done by conducting several in- depth interviews and document reviews. In this regard eight interviews with the relevant experts and managers of different projects have been done and some organizational documents have been reviewed.

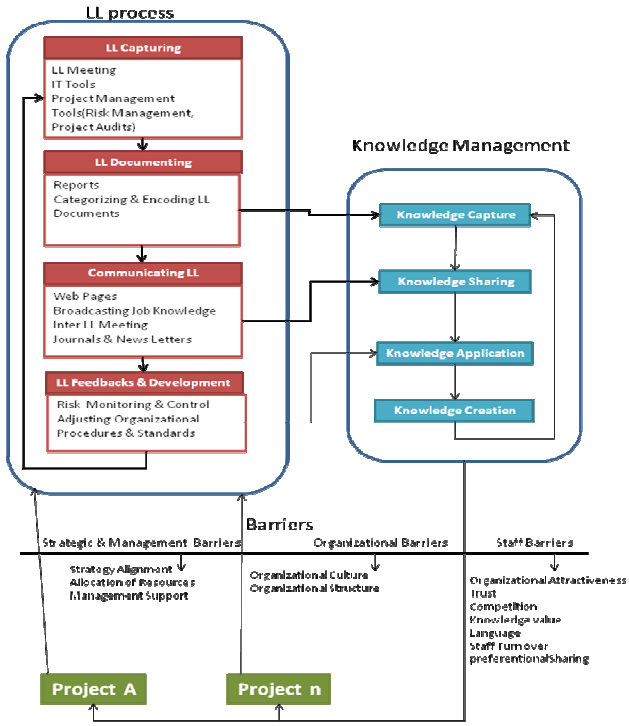


Figure 1: Conceptual Model of LL process and its barriers in project based organizations

In order to study the current status of the LL activities in projects, several questions related to the application of the LL in the organizations, their approaches toward LL as well as the reasons for negative attitude toward the LL were asked. These questions were: If the organization has any LL process. If the organization has any systematic approach for LL process. In case of negative answers, what is the reason? What are the problems and barriers to implementation of LL practices in the organization? The reasons are classified in categories of barriers to KM. Details of the findings are as below.

A. Strategic and Management Barriers

- Top Management Support- No Formal LL process: In most organizations, LL activities are ignored because they are not defined in the organizations' procedures and there is not a formal management support**.**
- Allocation of Resources for LL Processes: A quick look at the four steps of LL process presented in section 2.2 shows that implementation of LL process needs enormous effort including time and cost, which requires additional cost and expenses from the parent organizations.

B. Organizational Barriers

- Organizational Culture- Lack of Supportive Climate and Open Communication**:** As it is explained, supportive climate and managers'

 ISBN: 978-1-61208-181-6

power could create warm culture, trust, open communication system and fair rewarding system. These are important factors to implement a successful LL process. In the observed organizations these factors are not considered so much and leaders do not support and encourage people for collecting, sharing, and using lessons learned. Moreover, in this type of organizations, due to the political, accountability and organizational reasons, People may not like to participate in LL meetings.

C. *Communicational Barriers*

- Poor Knowledge Management & Communication Infrastructure**:** In most of the organizations there are no sign of integrated knowledge management information systems. In the exceptional cases, organizations have their isolated data, without any communications and exchange**s** of information between them. These deficiencies would be significant barrier**s** to implement LL process**.**

D. *Staff Barriers*

- Knowledge Value- Perceived importance of LL**.** Most of the senior and project managers do not know about potential benefits and real value of the systematic approach for LL. According the research findings, there are some traditional approaches in the LL activities, applied personally by some of the project managers, using the experiences of pervious projects. However. They do not know how much a systematic LL process can help to improve different aspects of their project and the parent organization's performance.
- Poor Attitude toward the Application of the LL: The common problem of the organizations is that there is not any hope and guarantees for application of the LL in other projects.
- Trust**:** From job security point of view, people prefer not to share their knowledge, they are afraid to lose their competitive advantages and their situation in the organization. Moreover**,** some managers believe that they are too busy to spend time looking back.

VI. PROPOSED SOLUTIONS

A. *Organizational Issues*

According to the case study findings, the most important problem**s** in our organizations are related to cultural issues. Below, there are solutions linked to different cultural factors.

- Open Communication Culture: Supportive climate and open communication can encourage people to share their knowledge and participate in LL activities [8]. A long period plan is needed to break the existing culture. The leaders have key roles to create trust (paired with staff barriers-trust) by a fair rewarding system and considering team building points as a major part of this strategy.
- Training & Learning Environment: Continual learning environment is the first principle for the LL process [13], this environment encourages people to share their knowledge and be active and creative in LL meetings [10]. Training is the first step to create learning environment. Lack of knowledge about Project Management and LL value (paired with staff barriers) can be solved by training. Theses programs also can help to remove the belief that LL process is costly or waste of time. When an organization decides to establish a formal LL process, some specific training courses can be helpful for better understanding of their tasks and implementing the process.
- Formal LL Process**:** With respect to the case study's findings, there is not any organizational obligation for LL activities, so establishing a formal LL process can be helpful. The process should be unified and consistence with all organizational sections. It should include codification and standardization activities, clarified tasks and responsibilities and considering incentive and feedbacks [22].

B. *Project Management Issues*

According to the case study's findings, below, there are the solutions linked to the project management techniques.

- Project Management Information System (PMIS): Developing an Integrated Project Management Information System (PMIS) has a significant role in development of LL & KM. In this regard, there are some recommended features in development of PMIS, from LL point of view, such as being user friendly and easy to use, having root cause analysis and searchable key word related to the LL data base. Additionally, having possibility for knowledge sharing and broadcasting of job knowledge which means producing a common language of unified data for dealing with problems from multiple locations and having a communication capability that automatically places information directly into the hands of the persons throughout the network of organizations.
- Risk Management Process: There are important interactions between three processes of Lessons learned, risk management, and communication management. Sharing knowledge in a systematic format, documenting LL, and ensuring frequent communication will maximize project success factors [4]. Risk management process can have a

 ISBN: 978-1-61208-181-6

significant role in the capturing and communicating project problems and pitfalls from LL point of view. Risk management helps to capture LL data as well as communicating them. On the other hand information related to the LL is used to risk identification and improve communication process. In addition project communication facilitates risk monitoring and control, and more communications about problems and risk factors would occur at LL meetings. Thus implementing these three processes can enhance each other during the project life cycle.

- Project Management Office (PMO): For development of the LL activities in the IOGP, there are some solutions, which are classified in organizational and managerial groups. These can be implemented by establishing of PMOs in the organizations. Establishing PMOs in organizations is a comprehensive solution that can include aforementioned solutions because the functionality of the PMOs can cover most of the requirements of the above solutions.

VII. CONCLUSION

In this paper, the main barriers of KM and LL in projects were investigated. Based on the research findings these can be classified in four major groups including strategic and managerial barriers, organizational barriers, communicational barriers and staff–related barriers. Investigation of the maturity and application of the LL in Iran Oil & Gas projects show that there is not a formal and systematic process for LL management. Moreover, there are many problems related to these barriers such as lack of supportive climate and open communication infrastructure, poor knowledge management & communication, poor attitude toward the application of the LL and perceived value of LL and trust. The proposed solution to improve the LL documentation consists of 1) organizational issues including creating open communication culture, training & learning environment and defining formal LL process 2) project management issues including developing project management information system (PMIS), risk management process and project management office (PMO).

REFERENCES

[1] C. Pritchard, Project Management Communication Toolkit, 1st ed., Artech House, Boston-London, 2004.

[2] PMI, Project Management Body of Knowledge (PMBOK), 4th ed., Project Management Institute, USA 2008.

[3] R. Wysocki and R. Mc Gary, Effective Project Management, 3rd ed., Wiley, New York, 2003.

[4] S. Seningen, "Learn the value of Lessons Learned" <http://www.projectperfect.com> 11.2011

[5] P. E. d. Love, P. S.W.Fang, and Z. Irani, Management of Knowledge in Project Environment, 1st ed., Elsevier Butterworth-Heineman, Oxford, 2005.

[6] B. P. Bergeron, Essentials of Knowledge Management, Wiley, Hobboken, Newjersey, 2003.

[7] R. Rynhardt, Barriers and Facilitators to Knowledge Management in Multi-National Companies: The Case of Nissan, University of Pretoria, 2008.

[8] D. Levi, Group Dynamic for Teams, 1st ed., Sage Publications, California, 2001.

[9] W. R. King, Knowledge Management and Organizational Learning (Annals of Information System), Dordecht Heidelberg, London, 2009.

[10] J. George and G. Jones, Contemporary Management, 4th ed., Mc Graw- Hill, New York, 2005.

[11] G. Probst, S. Rraub, and K. Romhardt, Managing Knowledge: Building Blocks for Success, Wiley, New York, 2000

[12] C. Lin Yu and K. Lin Lee, Critical Success Factors for Knowledge Management Studies in Construction, Department of Civil Engineering, National Taipei University of Technology, 2006

[13] G. Chin, Agile Project Management: How to Succeed In The Face of Changing Project Requirements, 1st ed., AMACOM, New York, 2004.

[14]C. Bobbe, "A New Process for Harnessing Past Experience." <http://www.chiefprojectofficer.com/article> 10.2006

[15] Ohio state University, "Wrap-up meeting." <http:// www.osu.edu/articles > 10.2006

[16] Trainers Direct, "LL Meeting Procedures." <http://www.trainersdirect.com/resources> 11.2011

[17] J. Meredith and JR. Mantel, Project a managerial Approach, 5th ed., Wiley, New York, 2003.

[18] A. Wiewiora, B. Trigunarsyah, G. Murphy, G. Gable, and Ch. Liang "The Impact of Unique Characteristics of Projects and Project-Based Organisations on Knowledge Transfer", Australian government's cooperative research centre program, 2009

[19] Th. H. Davenport, D. W. De long, and M. C. Beers, "Successful knowledge management projects", Sloan Management Review, pp. 43-57, winter 1998

[20] W. F. Boh, "Mechanisms for sharing knowledge in project-based organizations", Information and Organization, vol. 17, pp. 27–58, 2007

[21] J.J.J. Kasvi, M. Vartiainen, and M. Hailikari, "Managing knowledge and knowledge competences in projects and project organizations", International Journal of Project Management, vol. 21, pp. 571–582, 2003

[22] N. Meshkati, "Cultural Influence on the Implementation of Lessons Learned."<http://www.allbusiness.com/human-resources> 10.2006

 ISBN: 978-1-61208-181-6

Turnover and ICT Contribution in Organizational Knowledge Management

The case of employee turnover in portuguese real estate.

Filipe Fidalgo
School of Technology
Polytechnic Institute of Castelo Branco
Castelo Branco, Portugal
ffidalgo@ipcb.pt

Luis Borges Gouveia
Faculty of Science and Technology
University Fernando Pessoa
Oporto, Portugal
lmbg@ufp.edu.pt

***Abstract*— Organizations face a number of major transformations; one of the most important is that all have been suffering from growing employee turnover. This phenomenon makes organization loses not only potential but also customer relationships, image, routines, and other more subtle issues. In some cases, the most significant lost is clients trust with may have a direct impact on sales and profit and, also, perceived quality of service. For organizations where the business processes are less depending from machinery and heavily rely on human relationships, this problem is even more relevant, being real estate business one such good example. Additionally, if we consider the increase time needed to sell real estate propriety after recent subprime worldwide crisis, sometimes the broker who initiates the process is not the one who finalizes it. It is easy to retain in the organization information about activities we performed (the "what" and "when"; that we may collectively consider as explicit knowledge). However, we cannot say the same about the way those activities are performed (the "how"; that can consider more of tacit knowledge). To solve this situation, organizations must promote ways to retain tacit knowledge, in a way that it can be stored and disseminated through the organization. This paper discusses such issues taking into consideration real estate professionals, forms of action against this phenomenon. Assess the contribution of Information and Communication Technologies (ICT), formulating a conceptual model for the capture and knowledge transfer, using Grounded Theory to inform the model.**

Keywords-Knowledge Management; Tacit Knowledge; Explicit Knowledge; ICT; Turnover; Real Estate.

I. INTRODUCTION

Knowledge has always being a hot topic in organizations, but nowadays assumes a critical role, because of constant changes, fast decision cycles and a knowledge oriented economy. Drucker [1] already pointed this in 1988, based it on the following three points: the basis of employment changes from the office and manual workers to knowledge workers, who resist the model inherited command of military organizations; an economy that requires organizations to be innovative and entrepreneurial; and, finally (and according to the author the most important), an heavy use of information technology. Also, we can say that human capital is the most remunerative resource of any organization in the long run.

Current changes are more common and faster, with less time to react and even less to predict them. In the organizational perspective, "*everything has an increasingly tight lifetime*". Transactions change from local scale to regional scale, and from regional to international and global scale, becoming increasingly intense and less predictable, promoting additional levels of competition [2]. We live in a time that "*less is more and the time is now*" – creating a urge to act on moment. This reality is not exclusively in products but across the entire organization, processes, technology, and even people.

Considering human resources, we can observe a growing phenomenon: employee turnover. A few decades ago employment was considered as a relationship for life, both for employees and organizations. Nowadays, this relationship in most cases is very small in its time span. To analyze the workers flows in the Portuguese economy we use an administrative statistical source – *Quadro Pessoal* (QP) collected by the Ministry of Employment (MTSS). As we can see in Table I, more than half the working population is linked to an organization, less than 4 years.

TABLE I. WORKERS SENIORITY (IN YEARS) IN PORTUGUESE ORGANIZATIONS

Year	*Total Workers*	*Less than 1 year*	*1 to 4 years*	*5 to 9 Years*	*10 to 14 years*	*15 to 19 years*	*More than 20 years*
2007	*2967559*	*713897*	*883286*	*633051*	*272900*	*221758*	*242667*
		24%	*30%*	*21%*	*9%*	*7%*	*8%*
2008	*3016571*	*696045*	*954170*	*606046*	*294669*	*213196*	*252445*
		23%	*32%*	*20%*	*10%*	*7%*	*8%*

Source: [3], [4].

As a result of employee turnover, in many cases, organizations face a hiring process that is always time consuming and costly. Moreover, it is the loss of intellectual capital that those assets can represent. Furthermore, it is also needed to consider the time required for a new employee to be effectively productive. If in the hiring process organization has few to innovate, in the loss of intellectual capital organizations must create mechanisms to minimize it. Knowledge management can support the creation of such mechanisms. Organizations shall promote the capture and transfer of knowledge, so the impact of employee turnover does not represent the loss of organizational memory or, at least, minimize it. These issues assume particularly importance in organizations where the activities are not mostly made by machines, but by direct human contact. Real Estate is a good example of this kind of organizations, where the relationship between real estate agent (broker) and the

 ISBN: 978-1-61208-181-6

client (buyer or seller), depends most directly from the quality of the relationship between them. If we focus our attention analyzing seniority organizations data, considering the special case of real estate, we see that the values are even more significant (Table II).

TABLE II. REAL ESTATE WORKERS SENIORITY (IN YEARS) IN PORTUGUESE ORGANIZATIONS

Year	Total Workers	Less than 1 year	1 to 4 years	5 to 9 Years	10 to 14 years	15 to 19 years	More than 20 years
2007	21905	6135	8675	4223	1286	772	814
		28%	40%	19%	6%	4%	4%
2008	22539	5646	9496	4366	1485	704	842
		25%	42%	19%	7%	3%	4%

Source: [3], [4].

While the overall picture, up to 4 years of seniority in the company, had general values of around 55%, considering real estate in Portugal case these values are around 70%. The question is what organizations should (and can…) do to minimize the loss of organizational memory caused by this level of turnover. By organizational memory we mean the extension and amplification of knowledge as the key asset of a knowledge organization by capturing, organizing, disseminating, and reusing the knowledge created by its employees [5]. Dalkir also refers to the impact that of losing organization employees, by giving the example of NASA where "*60% of aerospace workers were slated to reach retirement age all within a few years of each other*", and by this, threatened the loss of valuable knowledge of the Apollo-era missions [5], in what can be called as employee generation problem turnover.

Not losing the traditional techniques like coaching and shadowing, the aim of this paper is to assess the contribution that information and communication technology (ICT) can add to this problem. This study reports the efforts that are conducted in the context of the first author doctoral program. As a result, the paper presents the problem and proposes a knowledge management approach to retain tacit knowledge in order to cope with employee turnover.

II. SET THE CONTEXT: A BRIEF LITERATURE REVIEW

A. Knowledge Management

As defended by several authors, the value of knowledge is from all the organizational assets, the most decisive in the production [1], [6], [7], [8]. Organizational changes were very deep and with high impact in recent years. On one hand, the opening of markets resulting from globalization produces many challenges and puts pressure on both times to react and adapt to evolving markets. On the other hand, the emergence of a knowledge-based society turns knowledge into more central organizational assets and ones that needed to be further understand and preserved.

Organizations are made of people and many are feeling that the knowledge of its human resources is its most valuable asset [9]. To succeed, a knowledge management initiative must have a robust theoretical foundation [5]. According to Dalkir, these models providing the widest possible perspective on KM Choo (1998), Weick (2001), Nonaka and Takeuchi (1995), Wiig (1993), Von Krogh and Roos (1995), Boisot (1998), Beer (1984), and Bennet and Bennet (2004). [5].

The Nonaka and Takeuchi Knowledge Creation Model have the major contribute to this project, so it will be present in more detail. The Nonaka and Takeuchi theory of organizational knowledge management (Fig. 1 illustrates the four modes of knowledge conversion that are the core of the overall knowledge-creation process) argues that knowledge creation is an ongoing process of socialization, explicit, combination and internalization [10]:

- Socialization: sharing individual tacit knowledge;
- Externalization: from tacit knowledge to explicit knowledge (codification of tacit knowledge in metaphors, analogies, figures and stories to create new concepts and then justify them before the corporate imperatives);
- Combination: in which the prototypes of new concepts are developed and incorporated into the organization;
- Internalization: this knowledge through learning by doing and experimenting, making tacit knowledge to be generated again.

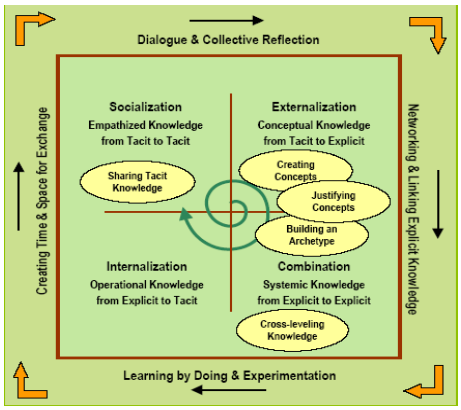


Figure 1. SECI Model [10].

The SECI or spiral model provides a good reference on how an organization deal with knowledge management issues and how a group of people are involved in the knowledge creation process. Knowledge creation always begins with the individual [5]. Making personal knowledge available to others in an organization is at the core of the Nonaka and Takeuchi Knowledge Spiral Model: this type of knowledge creation process takes place continuously and occurs at all levels of the organization – many times it occurs in an unexpected or unplanned way. Central to the SECI model proposal is the sharing of Tacit Knowledge that Davenport and Prusak define as complex knowledge developed and internalized by the professionals, over a long period of time which incorporates so much accrued and embedded learning that its rules may be impossible to separate from how an individual acts [8].

 ISBN: 978-1-61208-181-6

B. The Role of Information and Communication Technology

Traditionally, information and communication technology systems are used in organizations to support processes. One of current challenges is to make them support the professional competencies of individuals and to turn its adoption work in a broader collective context.

The use of computers and networks cannot stay only within the frame of operational tasks; they should add value to new forms of communication, conversation and learning on-the-job, support communities of practice, and provide the structure and access to ideas and experiences needed to excel in day-to-day organizational life. As stated by Davenport and Prusak, "*The computer's ability has little relevance to knowledge work, but the resources for communication and storage of networked computers make them enablers of knowledge*" [8] – this reinforces the role that human resources may have in knowledge organizations as the most value asset.

Without knowledge acquisition, knowledge transfer is meaningless. While knowledge transfer may be technology-enabled, knowledge acquisition is human-driven, so systems must be develop people-centered, not technology-centered [11]. Information and communication technology continues to be a powerful force in the ways in which people and organizations operate. ICT advances have become a permanent force bringing continuous and sometimes unpredictable changes to organizational structures and processes, including services delivered, management practices and governance [12].

Technological evolution allows tremendous freedom for creative thinking and a massive expansion of relationships, it also as a multiplier effect in promotion collaborative processes [13]. The use of computers, networks and digital information can open new opportunities in knowledge management and play important roles in meeting the prevailing challenges related to sharing, exchanging and disseminating knowledge.

As we know a large part of knowledge is not explicit but tacit, and this will a trend as human nature prevails as one of the most important sources for creativity. This is also true for knowledge in real estate business where a lot of good practices are transferred without being well documented in books, papers or any other documents. We must also notice that real estate business is largely based on face-to-face contacts and puts a real stress on human relationship. We defend that the use of ICT is needed to manage the knowledge properly.

C. Employee Turnover

A few years ago when we refer to employment, its general understanding that it was a lifetime relationship between the organization and the employee. Today this concept is passing by a radical transformation and must be redefined. A discussion about employment and its social role is presented by [14].

High turnover takes extreme importance, since human resource turnover represent costs to organizations. Chiavenato [2] proposes a cost list that are divided into primary, secondary and tertiary groups. The first group is quantitative, the second and third are qualitative estimates. Table 3 lists the Chiavenato costs for staff turnover [2].

TABLE III. COSTS OF STAFF TURNOVER

Primary	*Recruitment and selection costs;* *Registration and documentation costs;* *Integration costs;* *Separation costs.*
Secondary	*Production effects;* *Staff attitude effects;* *Extra labor cost;* *Extra operating cost.*
Tertiary	*Extra investment costs;* *Losses in business.*

Source: [2]

We may think, "*Some employee turnover is unavoidable, even desirable*". Some turnover is necessary, to replace some employees with more productive ones and to bring in people with new ideas and expertise. However, high turnover costs are both avoidable and unnecessary.

As new team members are added and others leave, it is critical to prevent the loss of information, even during such periods of major structural change [15]. Organizations can face a "Brain Drain" phenomenon if a turnover occurs and loses competent personnel at a higher rate than the organization can recruit and train new personnel. Some of costs involved to the organization or business could have easily been prevented in the first place if we experienced a way to retain employees´ knowledge. It is important to develop a strategy for retaining knowledge.

The knowledge lost from a departing employee is not a short-term problem; it is a long-term problem that breeds other problems and reduces an organization's effectiveness [11]. This is both a challenge and a problem that deserves to be dealt with.

D. Real Estate Bussiness

The real estate business has been in the last decades, one of the most important engines in western development economies. In first hand, if we live in modern cities, it is easy to observe the phenomena of empty places where nothing existed, but the real estate business changed them radically.

More recently, in 2008, the subprime crisis turns real state a more dangerous activity but still central for our economy. Nowadays we can expect to have a more challenging environment to buy and sells properties. In the context of Portugal, current sector statistics shows the importance of presenting smaller selling times, although there is a trend of the opposite.

1) Time to make real estate business

In the real estate business one key concept is selling time. Selling time can be defined as the amount of time needed for a certain property to be sold. It can be displayed in various units, usually days or months. For the independent Portuguese Real Estate Confidential (*Confidencial Imobiliário*) – which operates in two complementary areas

of business, editorial and production indicators of market analysis, the absorption time is the number of months that mediate between the first placed on offer and completion of the sale, understood as and the conclusion of the promissory contract of sale [16]. As for the ERA real estate network, presents itself under the name Average Days on Market [17].

Whatever the case, the reality is that this time has increased in recent years. According to data from *Confidencial Imobiliário*, in the last three years the increase was more moderate in the center of the country, more pronounced in the north of the country, while in the Algarve that time has nearly tripled [18]. The provided data by ERA real estate network in its most recent survey available, pointing in the same direction, although in this case the information is presented only in aggregate form, the values are global to the country and not by zone, as in the previous case. Indicating an average of 240 days on the market for 270 days for 2007 and 2008 [17].

Real estate organizations require real-time access to knowledge on a variety of subjects, including information on the core business and conditions affecting it, the business units' current objectives and corresponding real estate requirements, and the latest thinking in approaches to real estate [15].

Properties sales, like any complicated transaction, benefit from the attention and continuity of the real estate broker, and when that isn't possible, deal is delayed and can take more time to close because of staff turnover. New employees need time to establish relations with costumers, understand them, and then conclude the deal.

III. METHODOLOGY

We cannot look only consistency and validity in structured data, quantitative arrangements. At a time when the paradigm highlights the importance of the person as a guarantee of success – Knowledge Society – is increasing, the need for text analysis, interviews, speeches, among others, that is, pursue a qualitative analysis.

The qualitative methodological approach used in this research is the Grounded Theory. According to Fortin this "*aims to generate a theory from data collected in the field and among those who have relevant experience"* [19]. This theory began with Glaser and Strauss in 1967 and continued more expressive, with Corbin and Strauss in 1990.

Interview was used to collect data. This is indicated in cases where little knowledge exists about a particular phenomenon, and the researcher intends to obtain data on the research questions. In this particular case, partially structured interview were made. We interact through interviews with real estate professionals who hold a large experience, which gives them a deep knowledge of the subject under review.

Was defined a non-probabilistic sample selection rational, determined by the purpose of the study and theoretical relevance, and its potential for the development of the theory. The theoretical sampling aims not the representativeness of the sample, but the "*representation of the concepts*" [20].

Individuals are selected according to their level (expected), for generate new ideas for the elaboration of the theory [21].

The profile of respondents will be based on two different assumptions. On one hand professional experience in an area of more than 10 years, and on the other, organization seniority of at least 5 years. The first, it will ensure a thorough understanding of the study area, as well as all the elements needed for success in business. The second one gives us the double vision of the deep knowledge of the organization and the importance of the impact of turnover in the loss of skills. We contacted real estate agencies by email and when needed later by phone. Most of the interviews were conducted at the broker's agencies. In the interviews the researcher presents his study and the interviewed signed the informed consent – a document to authorize the data usage, as previously approved by the university ethics committee. This process took about 30 minutes per interview. A voice recorder device was used in the interviews so they later could be transcribed for analysis.

In the data analysis phase Strauss and Corbin proposed a method of comparative orientation composed of three types of coding [20]:

1) Open coding: are identified and coded, all statements, which the investigator deems significant, near or far the phenomenon under study;

2) Axial coding: preliminary codes will be compared and grouped according to their properties and dimensions and will allow the construction of conceptual categories;

3) Selective coding: categories intended to regroup for the construction of several major categories. These main categories will become central concepts and integrators to form the grounded theory to formulate.

In short, the researcher makes connections between all of the facts obtained, to construct a theory. To implement this methodology, first we select the "initial case", one significant case, associating with direct observation, in the environment or following the professional on the job and some relevant literature and we are in conditions make first analyses and to achieve the theoretical base (Fig. 2).

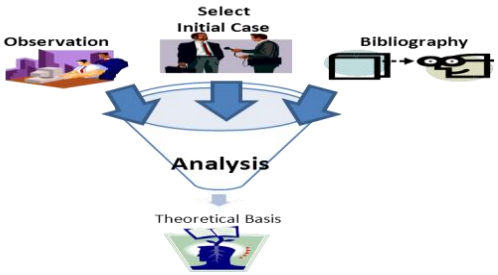


Figure 2. Grounded Theory Method – Theoretical Base.

After this first theoretical base formulation, followed an iterative phase where new cases are added (eg. other interviews) more observations and readings to make a new analyses (Fig. 3). If new findings appear, the theoretical base is reformulated and a new iteration begins; otherwise the theoretical saturation is achieved (when nothing new is found and the theory no longer suffers changes). Now the researcher formulates the theory.

 ISBN: 978-1-61208-181-6

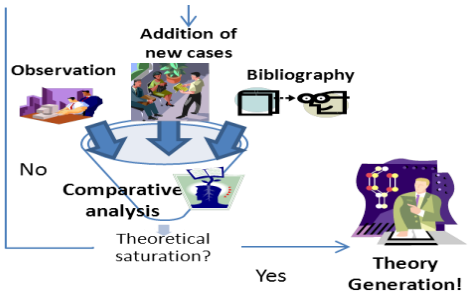


Figure 3. Grounded Theory Method – Theoretical Sturation.

In order to validate the theory an alignment must be made with the theoretical reference.

IV. RESULTS ANALYSIS

Although the main focus of this paper was to build a conceptual model linking Information and Communications Technologies, Employee Turnover to Knowledge Management, our results also provide an opportunity for some substantial comments on the Portugal Real Estate Business. Many efforts are made regard the use ICT, however, there appears to be a lag with respect to the advanced use of technologies such as video conferencing, intranets, etc. As we can see from one of the interviews:

"*...I can tell you that after two days I was about to leave because I did not realize anything about computers, and still don't, but ... now I perform a task in 10 minutes that at the beginning took me one day, two days and ended my patience*".

Another one refers:

"*... I don't say them aren't good, I don't know I to manage them*".

Other says:

"*...I'm old school; I'm more of the paper and pen time*".

However is common opinion that ICT are a fundamental resource on their jobs:

"*It's very, very important. Without Internet we can't do anything. When there is no system, we act like crazy*".

Employee turnover, and its effects in the organizations it's also a very concern issue:

"*...when a person leaves an organization, always exists a loss, because contact with the client was that person. And then when the next contact is another person, the client whether he will or will not retreat ...*".

"*...not only loss of knowledge/information, but also affect in a negative way by the loss of an asset created by the company. ...*".

Organizations also waste a lot of time settling down their new employees and make them productive:

"*...there were people, who learned quickly, and there were people that cost a bit more, it depends from one person to another. In some cases after two, three years had not yet heard a complete document or needed assistance to do it others after one month or two already knew how to do it ...*".

An indicator of this is that, for specialist positions responsible for KM, not a single position (i.e.: department, section) was found.

Results also suggest that many decision-makers still think that KM begins and ends with building sophisticated information technology systems and that no further organizational change is required.

Many others opinions were collected, but also point in the same way. In order to represent and show relations among these findings, a conceptual map was created. The theory about the concept map was developed in the 70s by Joseph Novak, he defines it as an administrative tool for organizing and representing knowledge [22].

There was made some relations between the phases from ground theory and the concepts representation in the concept map. Tree rings and a central concept are the present elements, as we can see in figure 4.

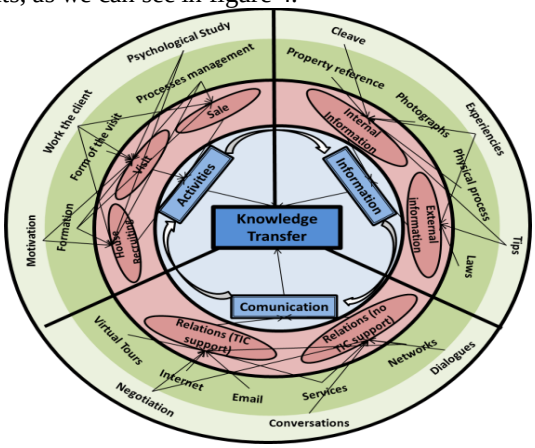


Figure 4. Conceptual Map.

The external ring, is divided in two parts, both represent concepts that we consider as atomic ones. A division exists between them. The light green (more exterior) is compose by elements from tacit dimension (eg. experiences, talk,…) and the dark green by elements from explicit dimension (eg. photography's, email,…). We can establish a relation between this ring and the open code from grounded theory. The second ring represents more aggregated concepts, generated by the first ones and his relationships, and in a similar way a relation with grounded theory axial code. In the last ring we found the principal concepts, as in grounded theory selective code, these main categories will become central concepts and integrators to formulate the theory. Real estate business has three major activities:

- House recruiting: real estate agent connect to the seller of a real estate property in order to represent him to the potential buyers;
- Visit: real estate agent show, showing the real estate property to the potential buyers;
- Sale: help the buyer and the seller of the real estate property making the transaction.

All this activities generate information that must be stored and disseminated throw organization. The information dissemination must be done using some communications channels. In fact what we are doing by performing these activities; stored the generate information and use communications channels to disseminated it, is achieve the main objective – knowledge transfer.

 ISBN: 978-1-61208-181-6

V. PROPOSED MODEL

In the former part, we have analyzed the results. In this section we propose a conceptual model (Fig. 5) in order to transfer tacit knowledge based in the concepts mentioned before.

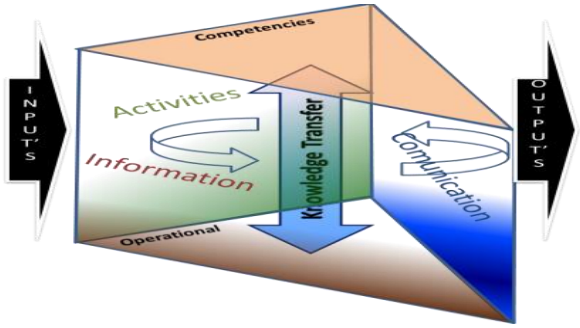


Figure 5. TATEK – Tacit to Explicit Knowledge.

A geometric figure was used – triangular prism to represent the model. The "*input's*" and "*output's*" arrows represents the model interaction with environment, what to receive and to provide.

Each lateral face of the prism represents the main categories mentioned in the conceptual map – Activities, Information and Communication. In all cases the colors nearby the inferior base are darkness and in the opposite side, nearby the top base are light. This difference represents the fact that either in Activities, Information and Communications we can found elements from two different dimensions of knowledge – explicit (dark ones) and tacit (light ones). Nearby the inferior base, the dark stands for the elements from explicit dimension. The inferior base represents the operational dimension, which give operational support to real estate business. In this dimension we can already found systems to support the explicit components from activities, information and communication. For example, when we schedule a meeting with a potential seller in order to recruit a new real estate property to represent we use ICT that give support to schedule the meeting, to store the result information and to communicate it to the organization.

But what if we think about the competences we needed? In all three dimensions Activities, Information and Communications, we use intangible elements. How can ICT give support in these cases? That's what the top base of the prism represent, an existing reality without support from ICT. We need to focus not only in creating systems that support operational task, but also systems to support competencies. Also create channels to disseminate them to the organization and make them available to operational systems. We can also make an alignment from the proposed model: TATEK, with the Knowledge Model from Nonaka & Takeuchi (Fig. 6). In the top base, previously identified as competencies dimension, we can find a correspondence with two elements of the knowledge conversion model from Nonaka & Takeuchi – Socialization and Externalization. In the TATEK model we assume that booth them can be supported by ICT. Nowadays, more people use ICT to socialize the Facebook platform is a good example. Emerging virtual worlds enable new ways to support knowledge and knowing processes because these virtual environments consider social aspects that are necessary for knowledge creating and knowledge sharing processes [23].

With the appropriate ICT tools workers can do the same in professional environment, creating reports of their success in the job tasks, additionally associating his reports with images, sounds, and video; and made it available to others in theirs organizations.

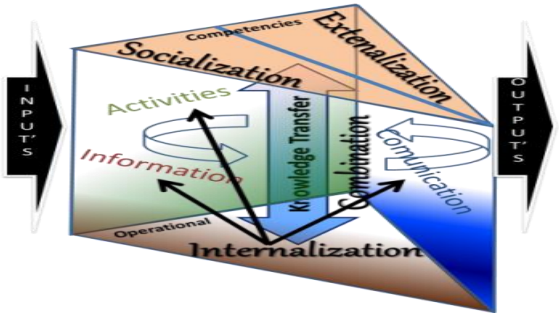


Figure 6. TATEK – Tacit to Explicit Knowledge - SECI.

Organizations play important roles creating conditions to make possible (and desirable) what we define as digital socialization and externalization, creating retribution programs to the knowledge providers is critical in this case.

These roles are also important to promote the combination of new knowledge's with the ones they already had. A successful KM implementation depends on a harmonious amalgamation of infrastructure and process capabilities, including technology, culture and organizational structure [24]. And, finally, evaluate the knowledge internalization, identified new activities, new information to provide and to receive to clients or to organizations and also new ways to communicate, otherwise in all three cases new approaches to the existing ones. This new knowledge represents the start point to a new cycle in the knowledge spiral.

Based on the *TATEK* model we can develop systems to share real estate stories (*SHREX – Sharing Real Estate eXperiencies*) – following a storytelling approach. This approach can take into consideration some clues to support its effectiveness like *raking*: scales showing the importance of each individual contribution and its organizational compensation (measures for Socialization and Extenalization); *ratios*: evaluating professional performance comparing who have and who have not access to the system (measures for Combination); *process modeling*: make a process model AS-IS comparing the before and after states and establish timeline making a new instance of the process model, allowing to identify innovations (in products, processes, activities, information, communication – measures for Internalization) .

VI. FINAL REMARKS

This research investigates the impact of employee turnover in organizational knowledge and influences of information and communication technology (ICT) in

 ISBN: 978-1-61208-181-6

knowledge management in the context of the Portuguese real estate organizations.

We use grounded theory methodology to collect data, associating direct observation, literature review and interviews to real estate agents in order to formulate a theory. Then, it was developed a conceptual model linking ICT to Knowledge Management.

Through our conceptual model, our researched, we are contributing to a better understanding of the principles and practices involved in building the foundations for Knowledge Management practice. Our findings about the relationships between ICT, Knowledge Management and Real Estate Business are relatively general. However, they prove to be particularly relevant for the Real Estate situation. On one hand Real Estate is a business depends particularly from relations between people, and consequently from the knowledge created by those contacts. On the other hand, the employee turnover has always been a phenomenon with which organizations have to confront. The point is that today, time to market pressure (which made it increasingly reduced) makes the increase turnover employee in Real Estate business a more operational and challenging problem.

As the time of a property stays in the market is rising, contributing to the central issue placed by our research, in many cases the real estate individual agent changes during the business. As a result, there is a need to create mechanisms to retain knowledge and allow continuity even with high rates of employee turnover.

The success and the competitive advantages of organizations came from the individual knowledge, so the ability to capture and disseminate it within the organization is a key factor for sustainable success. The abilities of an individual valued knowledge resulting from its activity (when speaking of human resources) should be retained in the organization so that their separation is not only an advantage to the host organization.

It is not enough to attempt to improve only one element, ICT, in its relationship with Knowledge Management, but it fails if not recognize Knowledge Management as essential for proper targeting Real Estate objectives. Progress depends on both technical and organizational change, and ICT professionals need to work closely with the others organizational players, in the deployment of Knowledge Management strategies.

The association of knowledge management and information and communication technology can leverage provide better systems that can support organizational success and cope with employee turnover.

REFERENCES

[1] Peter F. Druker, *O Advento da Nova Organização – Gestão do Conhecimento 13ª Ed. Harvard Business Review.* Rio de Janeiro: Editora Campus, 1998.

[2] Idalberto Chiavenato, *Administración de Recursos Humanos 5ºed*. Santa Fé de Bogotá: Editora MC Graw Hill, 2001.

[3] MSST. (2007) Web Site do Ministério do Trabalho e da Solidariedade Social. <retrieved:Set,2010> http://www.gep.mtss.gov.pt/estatistica/gerais/qp2007pub.pdf

[4] MSST. (2008) Web Site do Ministério do Trabalho e da Solidariedade Social. <retrieved:Set,2010> http://www.gep.mtss.gov.pt/estatistica/gerais/qp2008pub.pdf

[5] Kimiz Dalkir, *Knowledge Management in Theory and Practice.*: Elsevier. Butterworth Heinrmann., 2005.

[6] Thomas A Stewart, *Capital Intelectual – A nova riqueza das Organizações*.: Editora Silabo, 1999.

[7] Karl Erik Sveiby, *A Nova Riqueza das Organizações*. Rio de Janeiro: Editora Campus, 1998.

[8] Thomas H. Davenport and Laurence Prusak, *Conhecimento Empresarial – Como as organizações gerenciam o seu capital intelectual*, 8th ed.: Editora Campus, 1998.

[9] A. Serrano and C. Fialho, *Gestão do Conhecimento - O novo paradigma organizacional*.: Editora FCA, 2003.

[10] Nonaka Ikuhiro and Takeuchi Hirotaka, *Criação de Conhecimento na Empresa – Como as empresas Japonesas geram a dinâmica da inovação*, 20th ed.: Editora Campus, 1997.

[11] Hamilton Beazley, Jeremiah Boenisch, and David Harden, *Continuity Management: Preserving Corporate Knowledge and Productivity When Employees Leave*.: Ed. John Wiley & Sons., 2002.

[12] Jaro Berce, Sam Lanfranco, and Vasja Vehovar, *eGovernance: Information and Communication Technology, Knowledge Management and Learning Organisation Culture*. Eslovenia: Informatica an International Journal of Computing and Informatics., 2008.

[13] Sérgio Lins, *Transferindo Conhecimento Tácito – Uma abordagem construtivista*.. Brasil: E-papers Serviços Editoriais, 2003.

[14] Anthony Giddens, *Sociologia. 6ª Edição*. Lisboa: Fundação Calouste Gulbenkian, 2008.

[15] J. Samuells, *Putting knowledge management to work for real estate organizations*.: Real Estate Issues, 26(1):35–38, 2001.

[16] CI. (2006) Confidencial Imobiliário – Metodologia SIR. <retrieved:Nov,2010> http://www.confidencialimobiliario.com/sites/default/files/Metodologia_SIR.pdf

[17] ERA. (2009) ERA Europe Market Survey 2008/2009. <retrieved:Nov,2010> http://www.eraeurope.com/assets/pdf/ERAEuropeMarketSurvey_0809.pdf

[18] CI. Confidencial Imobiliário – Press Release. <retrieved:Nov,2010> http://ci-iberica.com/?q=content/press-release-tempo-de-absorcao-da-habitacao-no-algarve-triplica-em-3-anos

[19] Marie-Fabienne Fortin, *O Processo de investigação – da concepção à realização*, 5th ed.: Editora Lusociência., 2009.

[20] A. Corbin, J. Strauss, *Basics of qualitaty research*.: Sage Publications, 1990.

[21] Uwe Flick, *Métodos Qualititvos na Investigação Cientifica*.: Editora Monitor, 2005.

[22] J.D. Novak, *Concept Mapping: A Strategy for Organizing Knowledge*. Mahwah: Lawrence Erlbaum Associates, 1995.

[23] J Mueller, K Hutter, J Fueller, and K Matzler, "Virtual worlds as knowledge management platform - a practice-perspective.," *INFORMATION SYSTEMS JOURNAL*, vol. 21, no. 6, pp. 479-501, November 2011.

[24] Maria R. Lee and Yi-Chen Lan, "Toward a unified knowledge management model for SMEs," *Expert Systems with Applications*, vol. 38, no. 1, pp. 729-735, January 2011.

 ISBN: 978-1-61208-181-6

The Electronic Silverback

Absorb Knowledge Loss in Industry by Social Network Approach

Dirk Malzahn
OrgaTech GmbH
Lunen, Germany
dm@orgatech.org

Abstract—**Over the last decades, focusing on core competencies was one of the major management strategies to reduce cost and improve performance. Knowledge loss was accepted to some extent as this knowledge was not seen as crucial for the well-being of the company. What most companies underestimated was the impact of losing application knowledge, defining the specialties of applying non-core competencies on a company specific implementation. As both sides, the company and the supplier, limited their knowledge to their own core competencies, the required interfacing knowledge was completely lost. In this paper we will explain from an industrial perspective why this loss was disregarded for a long time, what the impact of this loss is and how the lost knowledge can be regained using a social network approach. Social networks are already widely used in industry, but mainly limited to marketing and recruiting. By this paper we want to extent the usage to the field of knowledge management. It builds the basis for a project starting in German Chemical Industry in 2012.**

Keywords- ***social network; industrial knowledge loss; core competencies; expert knowledge; incentives.***

I. INTRODUCTION

In the late decades of the last century, lots of companies reduced their processes and workforce to the minimum, required to deliver their so called core competencies. Under core competencies fall all elements that contribute directly to the creation of value of the produced product or delivered service, and build a unique selling proposition – or be at least not too easy reproduced by competitors.

All other competencies were outsourced to third parties, either by outsourcing parts of the company into independent companies or handing over work to a third party supplier. The basic idea was not too bad: let the company with the most experience and knowledge do the work it is specialized for - and by this benefit from their performance.

Companies easily accepted the loss in knowledge of outsourced work – as this knowledge was not seen as crucial and they still had suppliers to continue, and maybe optimize, this knowledge. By this approach, companies were able to free resources and money to further improve their core competencies and specialize on these - the same did the supplier. By this, another creeping knowledge loss started, which was not managed by most companies: the loss of application knowledge [1].

What is meant by application knowledge might best be explained by an example: if a company produces chemicals it needs pipes. Pipes are a highly standardized product. Therefore, the chemical company outsourced all piping work to an external supplier. Due to the high standardization of pipes, the supplier was able to deliver piping work in shorter time for lower cost.

With the saved money the chemical company was able to invest in more research for new products and optimized production processes. After several years of research the chemical company came up with a new product requiring a special type of piping. The company itself was not able to develop the piping system as it had outsourced all piping work. The piping supplier also was not able to deliver the piping work as it has focused on delivering standard piping systems and was afraid of the extra invest that has to be spend on research for developing the special piping system.

At the end of the last century, companies were still able to cope with this situation. Most companies still had employees "knowing the old times". In these old times, where everything was developed and delivered by the company itself, it was the employees' day-to-day work to cover all required steps of the production process. These employees usually were still able to give "hints" what a system should look like and therefore were the hidden knowledge tanks of the company – in one company they have been called "silverbacks" (like the gorillas) to express the deep technical knowledge they had collected during their lifetime.

By this, the problem was somehow known but not properly addressed as resources were still available to cover the problem.

But the longer it lasted, the more simple biological aspects came into account. Employees who collected their expert knowledge in the 70's or 80's of the last century are now retired – the silverbacks left the forest!

In Section II of this paper we will further define the problem. Section III describes the needs of the stakeholders. Section IV covers possible solution approaches. One of these approaches is detailed in Section V followed by a description of the implementation requirements in Sections VI to VIII. The paper ends with a conclusion, outlook and references in Sections IX to XI.

II. DEFINING THE PROBLEM

Focusing on core competencies leads to a knowledge gap where different knowledge areas come together. If both sides focus on their own knowledge area only, existing application

 ISBN: 978-1-61208-181-6

knowledge is stepwise lost. For some time, this can be absorbed by remaining part-knowledge on both sides and in later phases maybe also by knowledge silverbacks, but in the very end knowledge is irretrievably lost [2].

In the first phase, the company holds knowledge on the core competencies as well as on non-core competencies. By this, application knowledge is automatically maintained and developed.

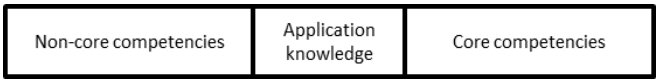


Figure 1. Starting state

Once the non-core competencies are outsourced, application knowledge decreases, the more both sides are focusing on their own core competencies.

Figure 2. Knowledge decrease

This lasts until the knowledge is limited to some experts.

Figure 3. Silverback state

In the very end, this leads to a knowledge gap.

Figure 4. End state - knowledge disconnect

Dependent on the state a company is already in, the problem has different severity. A company in starting state has to perform a knowledge management initiative to ensure that application management knowledge is collected, maintained and developed.

A company in knowledge decrease state has to do the same, but should also evaluate how much knowledge is already lost, e.g., by analyzing application knowledge need for possible research and development initiatives.

If the company is already in Silverback state it has to ensure that the knowledge of the Silverbacks is conserved and multiplied (e.g., by a mentoring approach, interviews, scenario techniques…).

The biggest challenge is to regain application knowledge for a company that is already in the knowledge disconnect state. Here once existing knowledge is lost and has to be re-build based on the requirements of the further developed core competencies.

The approach described in this paper covers the knowledge disconnect state, integrating elements from the Silverback state, as the first two states allow management by standard knowledge management techniques.

III. DEVELOP THE PROBLEM

To propose a solution, the impact, impediments and preferences of the different stakeholders have to be evaluated [6].

Stakeholders, in this case, are the company which is asking for the best solution, the supplier who is interested in developing a long-lasting customer relationship as well as the development of its own core competencies, and last not least the Silverbacks as long as they are available [3].

A. Company needs

The company first has to identify which application knowledge is required [4]. Based on this it has to be decided whether this application knowledge should be developed by the company itself or a supplier.

If the application knowledge is developed by the company itself, it has to cover research time as well as research invest. Based on time and invest it has to decide whether this application knowledge might or should become a core competency of the company.

If the application knowledge should be delivered by a supplier, the company has to cover the additional research cost and the risk, that the supplier is not able to deliver the application knowledge in the very end. If the company is developing the application knowledge itself, it is facing the same risk, but in this case the company is able to manage the risk directly, which is not possible if knowledge provision is outsourced.

B. Supplier needs

If a supplier develops application knowledge for a company, this work has to improve the suppliers' capabilities and economic success – otherwise there is no need for the supplier to perform this work.

Supplier capabilities are improved if the supplier can reuse the developed knowledge for other customers or improve its own core competencies.

Economic success is reached if either the cost for knowledge development is covered or the knowledge can be used for several customers and by this an economic benefit can be reached. Indirect economic benefit is delivered if the supplier becomes a preferred status and therefore is more often selected by the company.

C. Silverback needs

The Silverbacks are the most contributing but also the most problematic stakeholders. On one hand they hold most of the required knowledge. On the other hand they need a strong incentive to participate ("why should I help them on

 ISBN: 978-1-61208-181-6

something they have thrown away several years ago, now that I'm retired?").

IV. SOLUTION APPROACHES

As problem and stakeholders are known now, question is how the problem can be solved with the given stakeholders. Usually a formal approach is chosen.

A. Formal Approach

In a formal approach a project is set up to develop the knowledge. Therefore resources from company and supplier are required, which work on a defined topic to deliver defined results on a defined timeline.

The positives of a formal approach are that
- Structured management is possible
- Resources and roles are defined
- Goals are defined
- Effort and cost are planned

The negatives of a formal approach are that
- Required resources are hard to get (Silverbacks working for a defined duration with a defined effort)
- Effort and cost are hardly predictable if research methodology and impacting factors are unknown
- Project is limited to the goals, additional benefit opportunities identified during the project are not followed up

So, a formal approach will always help when we know what we need and want, and be able to define whom we need for this work, and can ensure that all the required resources are available.

B. Informal Approach

By an informal approach, information is collected in a community. Several people can work on a topic at their own will defining their own effort and contribution.

The positives of an informal approach are that
- A wide group can contribute
- Costs are minimized for the first step
- Additional benefit opportunities might be mentioned as well as impediments unknown by now

The negatives of a formal approach are that there is
- No guarantee that a solution will be developed at all
- No structured approach and timeline
- Unclear ownership of contributed knowledge

Given the management benefits of a formal approach and creativity benefits of the informal approach, combination of both might also be reasonable.

C. Combined Approach

In the combined approach, knowledge development starts with the informal approach. Driven by an event the collected information is transferred into the formal approach.

Possible events might be
- Information quality – amount of collected information and knowledge is sufficient to perform a reliable planning
- Time constraint – knowledge development project has to be started at a defined point in time to ensure in-time knowledge delivery. Up to this starting point as much information as possible is collected by the informal approach.

V. DEFINING THE INFORMAL SOLUTION APPROACH

Performing projects is a well-known and properly equipped process in industry [10; 11]. Therefore, the formal approach is not further described here.

Using the informal approach is much more uncommon in industry. Therefore, influencing factors and possible impediments require further analysis.

At first, the contributors (here stakeholders) act at free will. To foster this, each contributor must receive an incentive for the contribution.

Then, the contributors provide knowledge. This knowledge must be useful for other users. Therefore, the benefit of a contribution must be rateable.

As a third point, the user wants to get his problems solved, so there must be a possibility to communicate questions and problems.

Last, but not least, intellectual property must be safe. It has to be either ensured that only selected users have access to defined information or the incentive reaches a level that allows common usage.

VI. IMPLEMENTING THE INFORMAL SOLUTION APPROACH

The basic idea is to implement the informal solution approach by using the strategies and technologies known from social networks, in this specific case:
- Set up stakeholder specific profiles
- Connect to other stakeholders
- Share information (knowledge, questions, problems, etc.)
- Comment on shared information
- Reuse shared information
- Rate shared information
- Limit access to information
- Allow direct contact
- Calculate and deliver incentives
- Independent clearance

Each of the topics above will be further discussed now.

 ISBN: 978-1-61208-181-6

A. Setup stakeholder specific profiles

As described before, stakeholders are companies, suppliers and silverbacks. In the interaction, companies now become information customers of suppliers and Silverbacks. Customer companies want to get their problems and future challenges resolved. Therefore, it has to be possible to profile requests as well as fields of work. This allows other stakeholders to react on a problem or identify, develop or contribute knowledge useful for solving further challenges.

Suppliers mainly want to present themselves and show their capabilities. Therefore, there should be an appetite to present marketing information in addition to the competencies itself.

Major question regarding the Silverbacks is how it can be ensured that they are interested in participation. Thinkable might be a Silverback network which allows connecting to retired colleagues and a very strong incentive model ("what do I get out of it?"). On the other hand, it has to be ensured that required information about the Silverbacks' qualification and capabilities is available.

B. Connect to other stakeholders

Once stakeholders are present in the network, connecting types have to be evaluated. First question is whether direct competition is allowed. If so, each supplier can see the offers of its competitors and can react by pointing out which advantages it has compared to the competitors. Same is the case for customers. If they are able to see their competitors and also the suppliers working for them, they are able to benchmark themselves as well as to select new suppliers based on the contributions for other customers [9]. The direct approach allows building sub-communities (e.g., setting up a syndicate working together on a specific product or service for a defined period of time under defined rules).

On the other hand, it might be required to limit information to a specific group (e.g., if confidential information is made available). In this case the information offering partner should be able to decide which information is publicly shared and which is limited to a defined group.

Same is applicable for Silverbacks. They should be able to openly present themselves and their knowledge as well as limit access to private knowledge and conversations.

C. Share knowledge

Knowledge sharing is defined by profile and connection type. It should be possible to share structured knowledge as well as unstructured knowledge. Structured knowledge is offered based on a defined topic, question or problem. It can clearly be assigned to a specific field of work or application. Unstructured knowledge is each information, a supplier, customer or Silverback wants to offer. It has more a "what I also did in my life" style than addressing a specific topic.

Knowledge sharing should be able in a pull and push mode. In the pull mode [8], an interested party searches for information. In the push mode information is distributed to generally interested parties whenever it is produced. To avoid information overload, a subscribing mechanism must be in place to allow a pre-selection of acceptable "pushs".

D. Share questions and problems

Questions and problems are requests for information. Whilst a question is usually made available to all participants, a problem might require a proper pre-selection of involved parties. Background of a question is to retrieve as much information as possible on the specific topic. Problems look for more specified information and therefore have to avoid "information noise", meaning information which is generally useful but not contributing in resolving the problem.

It should be defined who is allowed to raise questions and problems. In a customer-centric network only customers should be able to raise questions and problems.

In a knowledge-centric network, everybody should be able to ask everybody else (e.g., Silverback asking for a specific tool, supplier asking for a special sub-process, etc.).

E. Comment on shared information

Once information is shared, everybody should be able to comment on it. Comments could be remarks, enhancements, corrections, but only the original provider should be able to change information based on the comments.

F. Reuse shared information

All information should be available to the intended user group. This can either be all participants or a limited group defined by the information provider (supplier makes new method available to preferred customer) or an information user (e.g., customer uses specific supplier knowledge).

Question is how an interested party gets hands on the information. Therefore manual and automatic search mechanisms must be available to identify and select required and useful information. Especially in research driven industries the "language" often is not fully defined, therefore search must be possible on syntactical and semantic level.

Additional search setting might be the current rating of information, to identify often used (common) information as well as seldom used (expert) information in relationship to the search topic. To avoid information noise, it has to be possible to exclude information rated as not useful.

Once information is selected, it has to be defined under which rules information can be re-used [5]. Silverback knowledge might e.g., be re-published by everyone as long as it is ensured that the originator receives his incentive for each reuse. Confidential, protected or trademarked information might only be reused under defined rules.

G. Rate shared information

Every information has to be rated on quantity, quality and domain level. Rating on quantity level means that it has to be measured, how often information has been accessed. Rating on quality means that each retriever of information

 ISBN: 978-1-61208-181-6

rates the benefit he gets out of the information. In the last step the search keywords leading to information are collected to improve the matching of search results.

Based on these 3 base ratings, a participants rating can be calculated. By this all participants can be grouped in classes like "contributors" (delivering a significant amount of useful information), "profiteers" (using more information than they are contributing), "blabbermouths" (contributing lots of information but with low benefit for others) or "freeloader" (which are using lots of information without contributing anything).

H. Limit access to information

As said before, it must be possible to limit access to information. Therefore an invite mechanism is required which allows a contributor to select participants who are allowed to retrieve this information. For search improvement reasons this might be structured stepwise. It might e.g., be possible to limit access to the information but not to related keywords. If a participant now searches for a keyword, he might see that information on this topic is available but still has to ask the information provider for access (e.g., access to trademarked / confidential information requiring a contract, non-disclosure agreement…).

I. Allow direct contact

To allow quick feedback or clarification, there must be a possibility to contact other participants directly. This could e.g., be done by chat functionality.

J. Calculate and deliver incentives

Incentives should be calculated based on the rating that a participant receives (as described under VI.G). The easiest way is to transform the different ratings into points. These points then might be changed into different incentives per group (Silverbacks, suppliers, customers). How these incentives might be managed is described in section VII.

K. Independent clearance

Even though the whole network is based on trust, misuse cannot be ruled out. For this reason an independent clearance body (e.g., trust center borne by all groups) should be installed.

Whenever a participant thinks that information is incorrect, copyright and related rights are not properly respected, or the ethics of the network are not followed, the participant can ask for clearance.

Additionally this clearance body can support the installation of the Electronic Silverback as described in section VIII.

VII. HOW TO MANAGE INCENTIVES

It is evident that incentives have to be managed differently per group: an incentive for a Silverback has to look different than an incentive for a supplier or customer [7].

A. Customer Incentives

For customers no specific incentives are required at first sight. Customers benefit from the provided information. However in later phases incentives might be reasonable, if customers act differently in information provision. If e.g., some customers provide information regularly ("contributor") whilst others only use information ("freeloader"), new Silverback or supplier information might e.g., be made available to contributors first, whilst freeloaders are granted access after a defined period (x days / weeks later).

B. Supplier Incentives

Suppliers might see the network as a possibility to tighten their relationship to their customers. On the other hand suppliers have to spend a reasonable amount of time and effort if they really want to contribute to the network. A possible incentive might be to integrate the supplier ratings into supplier performance monitoring. Most large companies have performance indicators for their suppliers. Contract terms and durations, supplier selection and preference and cost calculation are often driven by these indicators. If a supplier now has collected a high number of points in the network, this should improve the suppliers' performance indicator. For this reason it might be required to calculate customer specific ratings (e.g., not only "how often has information provided by this supplier been used" but "how often has information provided by this supplier been used by a specific customer"). These customer specific ratings should be confidential and might e.g., be managed by the clearance body.

C. Silverback Incentives

The hardest group to get is the Silverback community. What drives a Silverback to contribute time and knowledge? If the Silverback is still in the company, classic incentive models like salary increase might be possible as well as early retirement ("you can retire earlier if you are still available as expert from time to time").

Silverbacks already retired might be caught by receiving credit or direct or indirect payments.

Giving credit might e.g., been implemented by inviting highly pointed Silverbacks to conferences or setting up expert groups which meet in appropriate locations with high quality service.

Indirect payment might e.g., be done by providing Silverbacks with the newest computer technology, offering discounts in company owned stores or allowing them to participate in company rebate systems.

Direct payment might either result from points (points collected are converted in € or $), or improvement sharing (Silverback receives a percentage of the calculated

 ISBN: 978-1-61208-181-6

improvement benefit reached by information provided by him).

To ensure fairness of the calculations, this might also be performed by the clearance body.

VIII. THE ELECTRONIC SILVERBACK

Most of the activities described above look like delivering short term results. Customers ask questions or search help for problems. Silverbacks help on questions and problems and suppliers will mainly use it to combine marketing with customer relationship.

But, over time, the network will become a vault of knowledge. The more the network is used, the more information and knowledge will be available. This is the time to lift the treasure.

Based on the questions asked and searches performed, standard question and search strategies can be designed. Based on the information structure and keywords, information can be grouped and combined.

In the very end the network will become the Electronic Silverback: holding and improving application knowledge and establishing analysis and retrieval mechanisms to provide this knowledge in defined use cases.

IX. CONCLUSION

Companies are more and more using social networks. Whether it is Facebook, LinkedIn or Twitter, the number of companies represented there is increasing every day.

But the fields of application are still limited. Social networks are seen as marketing or recruiting platforms, but not to conserve knowledge. On the other hand everything required is already there – it just has to be used.

The major challenge for using a social network is trust in the network provider. Companies have to rely that information is not distributed unauthorized. Therefore the network provider has still to ensure an independent clearance body trusted by all partners.

X. OUTLOOK

In the current stage, it still has to be proven that a social network approach for companies delivers benefits beyond marketing and recruiting. The approach described in this paper offers another field of application.

Looking at the environment surrounding companies the way to social networking is irreversible: research is more and more performed in communities; innovative products are developed in open source communities; and even companies bring together experts from all over the world to improve their research, development and production capabilities.

So why not trying this in the field of knowledge management? This approach and the upcoming project should prove this, and lead the way from a selected group of customers, suppliers and Silverbacks to a repository for a whole sector or industry. Who is not willing to share might not be able to survive.

XI. REFERENCES

This paper does not have any direct references to other publications. Reason for this is that the paper is an independent industry approach not relying on any academic pre-work.

However basis of this paper are discussions with industry leaders on usage and possible benefits of social networks in industry. Major topics of these discussions were:

- How can we conserve expert knowledge?
- Can social media help?
- How do we get from expert knowledge to community knowledge?
- How do we get knowledge communities working in industry?
- How do we make this approach sustainable?

During discussions we often came to the point where one of the participants said “you know, I read the book from…” – indicating the basic thoughts on and perception of social media and collaboration usage in industry.

Knowing that the list below is neither complete nor comply with academic research requirements, it should give an overview, which topics already drive industrial decision makers.

[1] C. Anderson, “The long tail”, Hyperion, New York, 2006

[2] M. Gladwell, “The tipping point: How little things can make a big difference”, Hachette Book Group, USA, NY, 2000.

[3] J. Howe, “Crowsdsourcing. Why the power of the crowd is driving the future of business”, Crown Business, NY, 2008.

[4] J. Jarvis, “What would Google do?”, Harper Collins, UK, 2009.

[5] L. Lessig, “The future of ideas: The fate of the commons in a connected world”, Random House, 2001.

[6] R. Levine, C. Locke, D. Searls, and D. Weinberger, “The cluetrain manifesto”, www.cluetrain.com, retrieved 10.09.2011.

[7] C. Li and J. Bernoff, “Groundswell: Winning in a world transformed by social technologies”, Harvard Business Press, Boston Mass., 2008.

[8] L. Manovich, “The language of new media”, MIT Press, Cambridge, Mass., 2001.

[9] R. Scoble and S. Israel, “Naked conversations”, John Wiley & Sons, 2006.

[10] C. Shirky, “Here comes everybody. The power of organizing without organizing”, Penguin Books, New York, 2008.

[11] J. Surowiecki, “The wisdom of crowds”, London, Abacus, 2005.

[12] D. Tapscott and A.D. Williams, “Wikinomics. How mass collaboration changes everything”, Penguin USA, 2007.

 ISBN: 978-1-61208-181-6

Semantic search in a process-oriented knowledge base

Daniel Kimmig*, Andreas Schmidt†*, Klaus Bittner*, and Markus Dickerhof*
*Institute for Applied Computer Science
Karlsruhe Institute of Technology
Karlsruhe, Germany
Email: {daniel.kimmig, andreas.schmidt, klaus.bittner, markus.dickerhof}@kit.edu
†Department of Informatics and Business Information Systems,
University of Applied Sciences, Karlsruhe
Karlsruhe, Germany
Email: andreas.schmidt@hs-karlsruhe.de

***Abstract*—MinaBASE, a process-oriented knowledge management system, currently features a simple full-text search as well as a more sophisticated expert search. While the former is much easier to use, the latter has much higher precision and recall characteristics due to a more content-aware filter-mechanism. In this paper we present a combination of both approaches, which preserves the intuitiveness of the full-text search while offering the same precision and recall as the expert search. This is achieved by using a single input box for entering queries and semantically evaluating the given items according to the ontological concepts and relationships of the knowledge base and giving respective automatic suggestions. In contrast to regular autocompletion-widgets, the suggestions are not simple keywords, but rather elements of taxonomies as well as numeric input boxes for specification of their properties.**

***Keywords*-process knowledge management, microsystems technology, semantic search**

I. INTRODUCTION

At our institute we developed *MinaBASE*, which is a process-driven knowledge management system, that models manufacturing processes of microsystems technologies [1]. For information retrieval purposes an expert search has been implemented as part of it. This filter-mechanism offers several input boxes for the different types of properties of the information objects in *MinaBASE*, to define several interconnectable restrictions, which must be fulfilled by the entries of the result set. The input boxes support an automatic completion of values which are valid for the specific types of properties. The goal of this completion is to support the user while defining the restrictions. The restrictions are then mapped to database JOINs using a Criteria-API of a object relational mapping tool, with the result of a high precisioned search due to referential integrity. Since it is necessary to switch between the various input boxes with different meanings, it can easily become tedious to define multiple restrictions. The automatic completion is a helpful feature, but for their suggestion it is absolutely necessary for the user to know which values are possible for the respective type of property and therefore also which values can be contained in the knowledge base. While the filter-mechanism is useful for technical experts, inexperienced users are easily overwhelmed by the complex input elements. Admittedly you can model and approximately represent complex facts with such user interfaces, but the success of internet search engines like, e.g., Google or Bing has created a different level of scale regarding interactivity and usability concerning the input masks within the domain of information retrieval. This is why we implemented full-text search based on the popular Lucene [2] library, which allows a more intuitive interface. This requires the relationally structured data of the knowledge base to be transfered into a representation comprehensible by the search index. The indexing of the documents results in a very performant search, whose findability can even be increased using techniques from text mining such as tokenization, stemming, stopword removal, n-grams or synonymous expansion. Due to the shallow structure of the documents however, it is difficult to depict the ontological relationships, which can be found implicitly within the entities of *MinaBASE*. An example for this is the deduction of knowledge from a taxonomical classification of objects which are difficult to translate into the shallow structure of documents. This implicit knowledge is therefore not available during the search process, which results in a low precision and recall of the full-text search component. It is admittedly possible to complete this information by using hooks in the request handling of Lucene and performing repective database queries to augment the information of the search index, but then you nevertheless lose the performance advantage of optimized index structures. For this reason, there is a need for a different approach, which combines the ease-of-use of the full-text-search interface with the high precision and recall of the content-aware expert search. In this paper we'll present a variant of the search-component which is a combination of the previously described approaches of the current state of *MinaBASE*. The use of a central input box and the use of an intelligent mechanism which completes the field automatically for the support of the taxonomy relation of the technical aspects as well as their parameters are important issues.

 ISBN: 978-1-61208-181-6

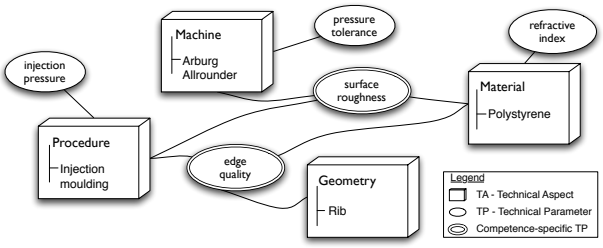


Figure 1. Schematic representation of a *MinaBASE* competence

The paper will be structured as follows: The next section will present the underlying process knowledge database *MinaBASE*. Then, a use case will be described that shows how the application-oriented search queries can look like. A solution approach will be described in Section IV, which will be evaluated briefly in Section V. The paper closes with a comparison of our approach with related work in Section VI and conclusions.

II. *MinaBASE* PROCESS KNOWLEDGE DATABASE

The micro system technology is considered to be one of the key technologies of the 21st century. It deals with micro systems, which means an intelligent, miniaturised system of sensors, data processing and/or actuators. Their product development is complex due to their small size, the free choice of materials as well as process-, design-, and manufacturing combinations. A specific manufacturing process is often required for each specific product, which results in low standardization of technologies. New challenges in the field of knowledge management are grown out of this, as there is a need to model manufacturing processes independent of specific products to facilitate product development by making it easier to retrieve universally valid knowledge about a manufacturing competence. An example for suchlike competences is the “injection moulding of PMMA on the Arburg Allrounder machine”, which is illustrated in Figure 1. The applied manufacturing procedure (injection moulding), the material (PMMA) and the machine (Arburg Allrounder) are central concepts of this competence. The so-called technical aspects (TA) that serve to model these materials, machines, and fabrication technologies are the smallest information entity in *MinaBASE* [3]. TA are arranged in taxonomies using generalization hierarchies. The number and contents of taxonomy trees can be specified and modified during runtime, such that a flexible structure tailored to microsystems technology can be defined. TA can be assigned properties that are referred to as technical parameters (TP). A TP is specified as a character string, integer, or floating-point number and references an attribute, e.g., density. As in the object-oriented approach, the TP of a TA are passed on to partial hierarchies located below in the taxonomy. In addition, lower hierarchy levels can further refine the inherited TP by specifying general value ranges. As a set of various TA from disjunct taxonomies, competences declare other TP, such as the edge quality and surface roughness [4].

III. USE CASE

After explaining the functional background, the structure of the underlying knowledge base we’ll now have a look at a use case, which shall demonstrate how a typical query can look like. For the comprehension of this use case it is important to know that the query doesn’t concern the details of the manufacturing competences from the technical point of view, but it is rather seen from an application-perspective. A first example for this is the use case ”cutting out a rectangle of a very thin, transparent, biocompatible plastic film” in context with *MinaBASE*, which is illustrated in the following Figure 2.

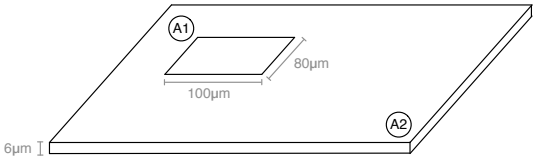


Figure 2. Structuring of a thin film polymer

It shows the physical properties of the desired application schematically. A rectangle of 100 μm width and 80 μm length shall be cut into a 6 μm thin, transparent and biocompatible plastic film. In this use case the rectangle, that is to be achieved, as well as the plastic film can be identified as the central input for the search, which can be conferred to the concepts geometry and materials, modeled in the aspect-taxonomies. The result of the search are those manufacturing competences, which are able to bring such forms in suchlike plastic films. The solution space can be narrowed down further by specifying more precise boundary conditions. One of these limiting conditions might be for example the exactness of the rectangle (quality of the edges or the precision of the sides angles). The search-component must be able to deduce these most relevant additional attributes appropriate to the use case by semantic interpretation of query terms and consultation of the knowledge base as well as offering corresponding context sensitive input mechanisms. Additional attributes for these use case are the dimensions of the rectangle (length and width) and also its precision. Due to this information you can make conclusions about convenient manufacturing procedures, which allow the creation of rectangles in a desired shape, like for example ”laser cutting”, ”laser milling” or ”precision milling”.

Furthermore the defined material qualities transparency and biocompatibility of the thin plastic form the starting basis for narrowing down the solution space to certain polymers, like for example polyurethanes (PU), polyethylene

 ISBN: 978-1-61208-181-6

(PE) and also polycarbonates (PC). With the aid of these basic information and the relationships which can be derived from the aspect-taxonomies (specialization of plastics, compliance of the additional attributes) a solution space of manufacturing competences can be constructed as a result of the query. Hereto the associated technical aspects need to be evaluated in order to find those competences, which describe the application of one of the mentioned process on the materials with the desired geometry and which correspond to the modeled additional attributes. In this manner an applied enquiry with qualities of the manufacturing competences can be seen as a solution for the use case. The actual possibilities regarding the query processing will be described in more detail within the next section.

IV. CONCEPT

Based on the previous use case, this section will explain the request handling for the semantic search and distingush cases of possible inputs. Figure 3 displays a mockup of user interface prototype. The requests will be initiated by entries into a single input box, which leads to a semantic interpretation of the query terms according to the concepts that are modelled in the knowledge base.

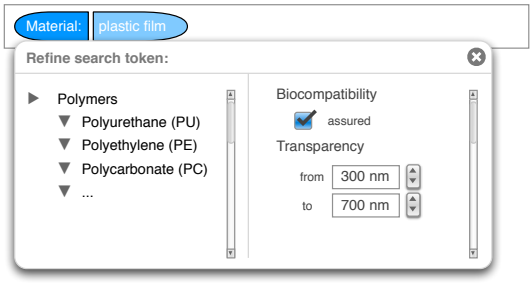


Figure 3. Semantic interpretation of keyword-based typed search tokens

As described, the central input box consists of a set of typed search tokens to which the system makes meaningful associated suggestions. The aim of this process is a better refinement of the search. The amount of such tokens can be seen as a set of restrictions, which need to be fulfilled by the competences which shall appear in the search result. Analogous to the described use case there is the plastic film as a base material filled into the input box in the Figure 3. Using synonyms of the material taxonomies, the engine suggests polymers and certain subtypes thereof. Additionally the engine finds most relevant parameters that are valid for the different taxonomy nodes like biocompatibility and transparency. After the user is done specifying the base material, he could for example enter the term ”rectangle”, which is the desired structure of the use case. In this case the engine would detect a node from the geometry taxonomy and load its most relevant parameters width and length as well as edge rounding and sidewall angle. While the user defines these restrictions, the result set of matching manufacturing competences, which are to be displayed underneath the suggestion box, is updated in real time in the background. This will enable an instant feedback mechanism enabling users to make incremental changes to their restrictions and see how these affect the result set. The rest of this chapter will focus on possible input values for the search-component and describe the specific cases in more detail. Figure 4 shows the possible interpretion of the given input. As you can see, one or more technical parameters (TA) can be specified. Additionally, the TAs can be described more detailed by adding constraints in form of technical parameters (TP) and even concrete values or value ranges for these TPs.

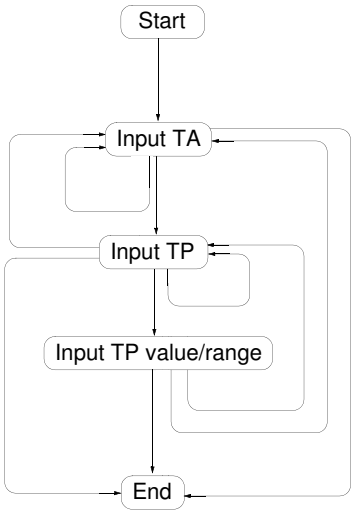


Figure 4. Input semantic for the keyword-based search

Case 1: Specification of a technical aspect

After specifying a technical aspect (for example a material), you can differentiate between the two following procedures:

(i) Overview on all technical aspects, which are below the selected node in the taxonomy: This makes sense, if you want for example to have an overview on the available technical aspects like the populated materials, processes, machines, etc. In this case the user interface offers all entries of the corresponding subtree. After the term is entered completely and recognized as a technical parameter these items are shown automatically by the suggestion-component and therefore can be selected directly. Afterwards the selected value overwrites the value that was entered before and represents a specialization of it.

 ISBN: 978-1-61208-181-6

(ii) Overview on the available technical parameters for this aspect: Every technical aspect is linked to one ore more technical parameters, which describe it more precisely. Furthermore these technical parameters are inherited from the root of the taxonomy to the leaf nodes.

To have an overview of the available technical parameters, all possible parameters for these technical aspect are shown in the input field after entering the technical aspect and the -tp (**t**echnical **p**arameter) flag. It contains the directly associated parameters as well as the parameters inherited from the upper levels of the taxonomy. By selecting one or more of these parameters and specifying a value or a range (see case 3), the search area within the taxonomy gets smaller as fewer nodes will fulfill the specified parameter value. Alternatively, instead of using the -tp flag, these parameters can be shown by focusing one of the entries in the subtree of the former case.

(iii) Overview of all technical parameters for this aspect and their specialization: If not only the parameters for the current technical aspect shall be shown, but also those parameters, which are defined below this aspect within the taxonomy, then this can be achieved by appending the flag -atp (**a**ll **t**echnical **p**arameters) after the technical aspect. If one of those parameters is selected, the origin of this parameter, which is a technical aspect that is located below the previous aspect in the taxonomy, is automatically selected and the previously typed or selected technical aspect is overwritten. This means that the selection of a specialized parameter will instantly lead to the selection of the technical aspect, that defines this parameter. Hereby a specialization of the previously entered term is achieved.

Case 2: Specification of a technical aspect and the corresponding parameter and items

After entering a technical aspect and a corresponding technical parameter, a concrete value or a range can be set for this parameter. If necessary the prior selected technical aspect gets replaced by an aspect which is in a lower part of the taxonomy, but only if it is required by the selected value or range of the parameter.

Case 3: Specification of more technical aspects

In this case the entry of a technical aspect (optionally with parameter and value/range) is followed by the entry of further technical aspects. Starting with the second technical aspect, the previously entered aspects functions as an additional filter. Therefore only those technical aspects are presented in the autocomplete box, for which there are competences, which also include the previously entered technical aspects.

Case 4: full-text search:

In case of not finding any terms within the concepts of the technical aspects, the technical parameters or the competences, the search-component switches automatically to the full-text search and then it tries to find hits within the foregoing concepts.

V. Evaluation

To evaluate our concept, we build a simple hardcoded version of our semantic fulltext flavored search. The first tests are encouraging, the search is much more precise as the previously implemented fulltext search. Additionally, after a short learning phase the time to retrieve the relevant datasets was significant shorter than using our old expert search.

VI. Related Work

The idea to bridge the gap between user friendly keyword-based search and expressive, formal and structured queries has fostered a lot of research in the past. The approaches that are closest to our paper are the following: SQAK [5] attempts to make it easier to access structured information stored in databases by using a keyword-based approach. Users require knowledge about SQL as well as a deep understanding about the structure of the underlying database for non-trivial use cases. SQAK attempts to overcome this situation by diagnosing the structure of the database and offering a keyword-based input mechanism requiring only small amount of knowledge about the structure of the database, that is translated into more complex SQL queries. DBXplorer [6] allows to search across relational databases using keywords by utilizing so called symbol tables that contain meta information about the database. Similiar to inverted document lists in information retrieval these symbol tables allow to match incoming keywords to rows and columns of the database, while also supporting queries spanning across multiple tables using dynamically built join trees. As we have only a limited set of tables (technical aspects and parameters) as starting points for our semantic search, the use of symbol tables is not adequate for our scenario. Our main focus is semantically evaluating the incoming search terms and suggesting, e.g., most relevant technical parameters or subconcepts for better refinement of the search restrictions. GINO [7] and Ginseng [8] provide a natural language interface for entering queries in a quasi-English language with guided entry of ontological concepts as well as derivation of triple-sets and SPARQL queries for query processing. While the concept of guided entry is similar to our automatic suggestion of parameters and aspects stemming from the knowledge base, we dont consider natural language interface suitable for our domain, since numeric input of parameter values and ranges is simpler to capture within a suggestion box since they are associated directly with the concept they are linked with in the knowledge base. XXploreKnow! [9], Q2Semantic [10] and SPARK [11] are focused on translating keyword queries into formal description logic queries. The approach of mapping search terms to knowledge base entities, exploring the connections between

 ISBN: 978-1-61208-181-6

them and utilizing the acquired knowledge thereof for further refinement of the search is similar to our approach. The difference is, the structure of their knowledge base is built on tools stemming from the field of the semantic web technologies, while *MinaBASE* uses a traditional relational database with a flexible entity-attribute-value schema.

VII. Conclusion and Future Work

This paper presented an approach to extending the *MinaBASE* process knowledge database, a system for managing the knowledge in the field of microsystems technology. By means of this approach, a semantic search can be implemented, that maps application-oriented properties to concepts of the knowledge base, which allows easier access and a more intuitive discoverability of knowledge entities. After the concepts of *MinaBASE* were explained in more detail in Section II, a concrete use case for the semantic search was shown in Section III. Afterwards we presented our concept in Section IV by first describing a mockup of the user interface according to the use case and then explaining the different input cases that need to be distinguished as part of query processing. In a future version, we plan to implement a customizable version of our semantic search. In this implementation the possible states, transitions, actions and preconditions associated with the transitions will be configurable. If this version will be implemented as a interpreter or with the help of a software generator is still an open question.

References

[1] M. Dickerhof, "Prozesswissensmanagement für die Mikrosystemtechnik." *Statusseminar MikroWebFab, Karlsruhe*, 2003.

[2] E. Hatcher and O. Gospodnetic, *Lucene in Action (In Action series)*. Greenwich, CT, USA: Manning Publications Co., 2004.

[3] M. Dickerhof and A. Parusel, "Bridging the Gap—from Process Related Documentation to an Integrated Process and Application Knowledge Management in Micro Systems Technology," *Micro-Assembly Technologies and Applications*, pp. 109–119, 2010.

[4] M. Dickerhof, O. Kusche, D. Kimmig, and A. Schmidt, "An ontology-based approach to supporting development and production of microsystems," *Proc. of the 4th Internat. Conf. on Web Information Systems and Technologies*, 2008.

[5] S. Tata and G. M. Lohman, "SQAK: doing more with keywords," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, ser. SIGMOD '08. New York, NY, USA: ACM, 2008, pp. 889–902.

[6] "DBXplorer: A System for Keyword-Based Search over Relational Databases," in *Proceedings of the 18th International Conference on Data Engineering*, ser. ICDE '02. Washington, DC, USA: IEEE Computer Society, 2002.

[7] A. Bernstein and E. Kaufmann, "GINO - A Guided Input Natural Language Ontology Editor," in *Proceedings of the 5th International Semantic Web Conference (ISWC 2006)*, Athens, Georgia (US), 2006.

[8] E. Kaufmann and A. Bernstein, "How useful are natural language interfaces to the semantic web for casual end-users?" in *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, ser. ISWC'07/ASWC'07, 2007, pp. 281–294.

[9] T. Tran, P. Cimiano, S. Rudolph, and R. Studer, "Ontology-based interpretation of keywords for semantic search," in *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, ser. ISWC'07/ASWC'07, 2007, pp. 523–536.

[10] H. Wang, K. Zhang, Q. Liu, T. Tran, and Y. Yu, "Q2Semantic: A Lightweight Keyword Interface to Semantic Search," in *In Proceedings of the 5th International Semantic Web Conference (ESWC'08*, 2008.

[11] Q. Zhou, C. Wang, M. Xiong, H. Wang, and Y. Yu, "SPARK: adapting keyword query to semantic search," in *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, ser. ISWC'07/ASWC'07, 2007, pp. 694–707.

 ISBN: 978-1-61208-181-6

Strategic Market Analysis in an Electronic Service Market

Gülfem Işıklar Alptekin

Galatasaray University
Department of Computer Engineering
Istanbul, Turkey
gisiklar@gsu.edu.tr

***Abstract*- The electronic-book (e-book) is one of the new technological changes that have significantly influenced the publishing industry in the last century. This has forced publishers to reconsider their distribution channels, since the Internet has provided a new means with which to serve readers. In this paper, a strategic market analysis is proposed from the perspective of a traditional publisher that needs to decide whether to switch to e-publishing business. The analysis framework determines the publishing market equilibrium in three different market scenarios. Besides, it shows the impact of readers' choices and price sensitivities on the profits of publishers. The proposed framework has its basis on game theory and it is built in an oligopoly setting to reflect the severe market competition. The readers' utilities and demands are modeled using the multinomial logit model. Although the first scenario possesses a global optimum solution, in the remaining two market scenarios genetic algorithms are used in order to find the sub-optimal solutions of the oligopolies.**

Keywords- distribution channel; multinomial logit model (MNL); game theory; genetic algorithm

I. INTRODUCTION

Traditionally, the term "publishing" refers to the distribution of printed works such as books and newspapers. With the advent of digital information systems and the Internet, the scope of publishing has expanded to include electronic resources, such as the electronic versions of books and periodicals, as well as micropublishing, websites, blogs or video games. In this work, an electronic book (e-book) will be considered as the digital version of a traditional printed book (p-book) to be read digitally on a personal computer, handheld computer, PDA or a dedicated e-book reader. It has been estimated that e-book sales will account for 50% of the publishing industry's sales by 2020 and 90% by 2030 [1].

In this paper, a strategic market analysis framework in a publishing market is proposed in the presence of multiple competing publishers. The proposed publishing market consists of p-publishers (publishers of p-books) that try to decide on whether or not switch to e-publishing (publishing e-books). In this environment, a p-publisher is assumed to face three market scenarios, presented in sub-section 2.2. In each scenario, the publishers' decision variables are their offered unit prices. There are two types of prices for an e-publisher: A price for its traditional retail channel and a price for its electronic/Internet channel, while there is only the traditional retail price for a p-publisher. The traditional channel and Internet channel of a publisher are distinguished based on two features: the stocking and maintenance costs and price sensitivity of readers. It is assumed that readers will be more sensitive towards price, when they intend to buy a book from an electronic channel.

The proposed framework computes the unit prices, and accordingly the profits of the publishers in each market scenario. Since the publishers need to make their decision in an uncertain market environment, the customer utility and demand are modeled using the multinomial logit model (MNL), which is based on probabilistic theory, by assuming that a customer has also a no-purchase alternative. For each scenario, a non-cooperative pricing game is built, whose players are the publishers. Solving the game, the mutual best response strategies that determine the equilibrium point(s) are studied. With the optimum prices, the publishers calculate their expected demands and expected profits. For the first scenario, the equilibrium prices and profits are global optimum, since the profit maximization problem is proved to be convex. However, the maximization problems of the second and the third scenarios cannot be proven to be convexes, hence their optimum solutions are found using a type of local search algorithm: the genetic algorithms (GA). A genetic algorithm is an evolutionary optimization approach which is most appropriate for complex non-linear models where finding the location of the global optimum is a difficult task [2]. GA utilize a population of solutions in the search, rather than handling one feasible solution like in other local search methods, such as Simulated Annealing or Tabu Search. GA have good performance in large and complex search spaces, since they explore and exploit simultaneously the search space. They do not guarantee global optimality even it may be reached.

The rest of the paper is organized as follows. Section 2 discusses the related work in the literature and their differences from this one. In Section 3, the formulation of the model, three possible market scenarios and their demonstrative examples are given in detail. The results are discussed at the end of the Section 3. Finally, conclusions and future work are given in Section 4.

 ISBN: 978-1-61208-181-6

II. RELATED WORK

In literature, the GA have been applied frequently as a part of decision support systems. In one of the recent works, the authors have used genetic algorithms in identifying the optimal parameters for water resource modeling applications. Moreover, they have optimized the genetic algorithm model parameters [3]. In literature, it is possible to encounter the use of GA into the game theory-based models. These works are from various research areas. Riechmann [4] has shown that economic learning via GA can be described as a specific form of an evolutionary game. In his paper, he has pointed out that GA learning results in a series of near NE. In another research, the authors have discussed a new evolutionary strategy for the multiple objective design optimization of internal aerodynamic shape [5]. They have claimed that game theory replaces a global optimization problem by a non-cooperative game based on Nash equilibrium with several players solving local constrained sub-optimization tasks. The authors have stated that game theory is not only the primary method for the formal modeling of interactions between individual, but it also underlies how biologists think about social interactions on an intuitive level [6]. In their work, they have used GA as an alternative method of searching evolutionary stable sets in a well-studied game of biological communication. In a sense, the point of view in this work resembles to the one in the proposed framework in this paper, since GA are used to search the market equilibrium points in games for three different market scenarios.

There are numerous research in the literature on the adoption of e-books and e-publishing; however this paper's concern is a more specific area. The concentration is on the economics and management aspects of e-publishing. The most relevant research are as follows: Jiang and Katsamakas examine how the entry of an e-book seller affects strategic interaction in the book markets and impacts sellers or consumers [7]. Their work is a good example of the application of game theory in analyzing the market asymmetries. The research of Hua et al., has the same research question as the one in this paper [1]. In their research, the authors derive the conditions under which a publisher should sell only p-books, only e-books, and both of them simultaneously. They use the newsvendor model to analyze demand behavior; whereas the demand is modeled using MNL model in this paper. As the demand varies linearly on offered price, they can determine the closed-form expression of optimum prices. In this paper, the problems are solved using nonlinear techniques. Bernstein et al. use the MNL model for the equilibrium analysis of retailers [8]. They differentiate retailers' choices as "bricks-and-mortar" and "clicks-and-mortar", which represent the traditional retail channel and Internet channel, respectively. Their study has some common grounds with the one in this paper, since they analyze the supply chain channel structure choice in an oligopoly setting.

III. THE MODEL FORMULATION

An *n*-firm oligopoly setting is considered to study the structure of the game [9]. In the proposed game, the publishers that sell their books through a retail store (*p-publishers*) want to reach more reader by publishing their books on an Internet channel (*e-publishers*). The *e-publisher*$_i$ will continue to sell its books on the retail stores, hence it has to define two different prices: a unit retail price p_i for its traditional channel and a unit online price p_{ei} for its Internet channel. The *p-publisher*$_i$ needs to define only its unit retail price p_i. As the stocking and maintenance costs of an e-book are assumed to be lower than the ones of a traditional book, the following assumptions on the prices are set: $p_{ei} \le p_i$. $A = \{1, 2, \ldots, n\} \cup A_0$ denotes the set of publishers.

A. Customer Utility Model

A reader is assumed to derive a different utility when obtaining the book from a retailer's physical store (alternative *i*) than obtaining it in an electronic form (alternative *ei*). Furthermore, a reader is assumed to have a no-purchase alternative (A_0). In other words, if s/ he does not like any offer, s/ he will not buy any book. Then, the set of alternatives is $A^P = \{1, 2, \ldots, n\} \cup A_0$ when all the publishers sell from their retail stores, while it is $A^E = \{1, e1, 2, e2, \ldots, n, en\} \cup A_0$ when all publishers sell both from their retail stores and their online stores. It is also possible to have a case with *k* e-publishers and (*n*-*k*) p-publishers, then the set of alternatives is $A^{EP}(k) = \{1, e1, \ldots, k, ek, k+1, \ldots, n\}$.

The customer utility is modeled using the multinomial-logit model (MNL). The MNL model is one of the random-utility models that are based on a probabilistic model of individual customer utility [10]. Let us assume that a customer has a utility for alternative *i*, denoted U_i. The probability that a customer selects alternative *i* from a subset *A* of alternatives is given by:

$$P_i(A) = P\left(U_i \ge \max\left\{U_j : j \in A\right\}\right) \qquad (1)$$

If we assume that $u_1 = -bp$, this gives the following demand function:

$$d(p) = M \frac{e^{-bp}}{1+e^{-bp}} \qquad (2)$$

where *M* is the market size and *b* is a coefficient of the price sensitivity. In the multiple-product case, by considering each user of the same type (b_i=b), the demand function of *publisher*$_i$ is given by:

$$d_i(p_1, p_2, ..., p_n) = M \frac{e^{-bp_i}}{1+\sum_{j=1}^{n} e^{-bp_j}},\ i = 1, ..., n \qquad (3)$$

The MNL probability that a customer chooses product *j* as a function of the vector of prices p = $\{p_1, p_2, .., p_n\}$ is then given by:

$$Prob_j(\mathrm{p}) = \frac{e^{-bp_j}}{1+\sum_{i=1}^{n} e^{-bp_i}} \quad (4)$$

B. Choice of Channel Structure

In the proposed strategic analysis, each p-publisher that has an incentive to publish its books on an electronic environment, in other words that has an incentive to move to e-publishing business, faces three marketing scenarios:

1. *P-P competition*: All publishers in the market are p-publishers.
2. *E-P competition*: Some publishers remain as p-publishers, but the rest moves to e-publishing.
3. *E-E competition:* All publishers in the market are e-publishers.

The list of notation for the equilibrium analysis under different scenarios can be given as:

p_i^{PP} : Equilibrium price of *p-publisher*$_i$ under P-P competition,

p_i^{EP} : Equilibrium price of *p-publisher*$_i$ under E-P competition,

p_{ei}^{EP} : Equilibrium price of *e-publisher*$_i$ under E-P competition,

p_i^{EE} : Equilibrium price of *p-publisher*$_i$ under E-E competition,

p_{ei}^{EE} : Equilibrium price of *e-publisher*$_i$ under E-E competition,

$\prod_i^{PP}$: Equilibrium profit for *p-publisher*$_i$ under P-P competition,

$\prod_i^{EP}$: Equilibrium profit for *p-publisher*$_i$ under E-P competition,

$\prod_{ei}^{EP}$: Equilibrium profit for *e-publisher*$_i$ under E-P competition,

$\prod_{ei}^{EE}$: Equilibrium profit for *e-publisher*$_i$ under E-E competition.

1) Scenario I: P-publisher vs. p-publisher (P-P Competition)

In this scenario, the set of alternatives for readers is $A^P = \{1,2,...,n\} \cup A_0$. All p-publishers simultaneously set their prices. Each p-publisher's aim is to define its optimum price (p_i^{PP}) in the given market environment that maximizes its profit:

$$\max_{p_i^{PP}} \left(\prod_i^{PP}\right) = \max_{p_i^{PP}} \left(p_i^{PP} - c_i\right)\left(M.\frac{e^{-b.p_i^{PP}}}{1+\sum_{t=1}^{n} e^{-b.p_t^{PP}}}\right) \quad (5)$$

$$\text{s.t. } p_i^{PP} - c_i \geq 0, \quad \forall i$$

$$p_i^{PP} \geq 0, \quad \forall i$$

where b is the price elasticity and c_i is the unit cost of *p-publisher*$_i$. The probability that *p-publisher*$_i$ with the price p_i^{PP} is chosen by a customer is given as:

$$\mathrm{Prob}_i^{PP} = \frac{e^{-b.p_i^{PP}}}{1+\sum_{t=1}^{n} e^{-b.p_t^{PP}}} \quad (6)$$

The first order condition of the choice probability with respect to its price ($\partial \mathrm{Prob}_i^{PP} / \partial p_i^{PP}$) is negative, which means that the price increase of a *p-publisher*$_i$ reduces its own demand; whereas the first order condition with respect to its competitor's price ($\partial \mathrm{Prob}_i^{PP} / \partial p_t^{PP}$), $i \neq t$ is positive, which means that the price increase of the competitor's price increase the demand of *p-publisher*$_i$. As p-publishers determine their prices simultaneously, they need to consider the competition, i.e., the prices offered by other p-publishers in their market. The problem is modeled as a game where the *players* are the p-publishers, the *strategies of the players* are their offered unit prices and the *payoffs of the players* are their profit functions. Solving such a game means predicting the strategy of the publisher. One can see that if the strategies from the players are mutual best responses to each other, no player would have to deviate from the given strategies and the game would reach a steady state. Such a point is called the Nash equilibrium (NE) point of the game [11]. In the game, p-publishers determine their prices independently and the information is strictly limited to local information. Hence, the game has a non-cooperative setup. Global optimality conditions are used in order to analyze the existence and the uniqueness of the equilibrium point. The constraints in the proposed problem (5) are linear, so they are convexes. Therefore, the vector $\mathrm{p} = \left[p_1^{PP}, p_2^{PP}, .., p_n^{PP}\right]$ denotes the solution (the NE) of this game with: $p_i^{PP} = BR_i\left(\mathrm{p}_{-i}^{PP}\right)$, where p_{-i}^{PP} represents the vector of best responses of all *p-publisher*$_t$, $t \neq i$. The NE is the point that solves the set of equations: $\frac{\partial \prod_i^{PP}}{\partial p_i^{PP}} = 0, \forall i$. It is also the global optimum of the given problem.

This scenario is demonstrated on a simple but representative example with two p-publishers in the market. The target customer group is assumed to be consist of M=100 readers. The unit costs of two p-publishers *(c_1=c_2)*

are assumed to be the same and equal to 1. The price sensitivities in the readers' demand functions are differentiated in order to analyze the impact of readers' price sensitivities on the publishers' price determination. The price sensitivity of customers of p-publisher$_1$ (b_1) is taken as 1, whereas the one of customers of p-publisher$_2$ (b_2) is taken as 1.5. For the first scenario, the equilibrium price values and related demands and profits are given in Table 1. As the price sensitivity of the customers of *p-publisher$_2$* is set higher than the one of *p-publisher$_1$*, *p-publisher$_2$* offers a lower price (p_2^{PP}= 1.7123) in the equilibrium. The demand values in given tables are found using Eq. (3).

TABLE I. PRICE, DEMAND AND PROFIT VALUES AT EQUILIBRIUM IN SCENARIO 1

	Equilibrium values	
	p-publisher$_1$	***p-publisher$_2$***
Offered price	2.1123	1.7123
Profit	11.2342	4.5592
Demand	9.2907	13.8630
Total profit of p-publishers	15.7934	

2) Scenario II: E-publisher vs. p-publisher (E-P Competition)

In the second scenario, the first $k\left(1 \le k \le n\right)$ publishers are assumed to be move on e-publishing, whereas the remaining *n-k* publishers are stayed as p-publisher. The set of alternatives for consumers is $A^{EP}(k) = A^{P} \cup \{e1, e2, ..., ek\}$. In this scenario, *p-publisher$_i$* determines only one price (p_i^{EP}), but *e-publisher$_i$* determines both a price for its traditional channel (p_i^{EP}) and a price for its Internet channel (p_{ei}^{EP}). The p-publisher's price is assumed to be influenced from other p-publisher's prices, whereas the e-publisher's price is influenced from both other e-publishers' prices and from the price of its own traditional channel. In other words, if *e-publisher$_i$* increases its Internet channel price (p_{ei}^{EP}), the demand to its traditional channel increases. From this point of view, two channels of an e-publisher can be considered as "competing" [8]. Both type of publishers' aim is to define their optimum prices (p_i^{EP} and p_{ei}^{EP}) that maximize their profits. *P-publisher$_j$* wants to maximize its profit:

$$\max_{p_j^{EP}} \left(\Pi_j^{EP}\right) = \max_{p_j^{EP}} \left(p_j^{EP} - c_j\right) \left(M . \frac{e^{-b.p_j^{EP}}}{1 + \sum_{t=k+1}^{n} e^{-b.p_t^{EP}}} \right)$$

$$\text{s.t. } p_j^{EP} - c_j \ge 0,\ j = k+1, k+2, ..., n \qquad (7)$$

$$p_j^{EP} \ge 0,\ j = k+1, k+2, ..., n$$

On the other hand, the objective of *e-publisher$_i$* is to choose p_i^{EP} and p_{ei}^{EP} that maximizes its own profit: The first term of the profit function belongs to the profit earned from p-publishing, while the second term belongs to the profit earned from e-publishing. In the proposed model, the coefficient $\vartheta \ge 1$ is inserted to the demand function because an e-reader's sensitivity to price is assumed to be higher than the one of a p-reader. In this setting, it is not possible to derive closed-form expressions for the equilibrium prices, demands and profits. The convexity of the maximization problem in this scenario cannot be demonstrated. Therefore, the GA is used as a computing technique to find sub-optimal solutions. In the GA implementation, three different population sizes are utilized: 100, 125, 150. The profit functions of the publishers are used as fitness functions in order to decide the chromosome in the next generation. For the GA, the most frequently used stopping criterion is the specification of a maximum number of generations. In the GA implementation of the scenario, the maximum number of generations is defined as 250, which means that the algorithm terminates once the iteration number reaches 250. For each population size, the GA is run 50 times and the best result is chosen from these 50 results. The solutions have shown that the optimum solution does not depend on the initial population size in this scenario. The second scenario is demonstrated on an example with *e-publisher$_1$* and *p-publisher$_2$* in the market. For the second scenario, the equilibrium price values and related demands and profits are given in Table 2. The results at the equilibrium point confirm that the e-publishing price is lower than the p-publishing prices because of lower publishing costs. *E-publisher$_1$* reaches bigger market share, which is proportional to its total profit, since it offers two different publishing channels for different preferences.

TABLE II. PRICE, DEMAND AND PROFIT VALUES AT EQUILIBRIUM IN SCENARIO 2

	Equilibrium values		
	e-publisher$_1$		***p-publisher$_2$***
	e-publishing	***p-publishing***	***p-publishing***
Offered price	1.3839	1.7124	2.1426
Profit	14.0419		4.5729
Demand	9.6630	13.1912	9.9370
Total profit of publishers	18.6148		

3) Scenario III: E-publisher vs. e-publisher (E-E Competition)

In the last scenario, it is assumed that all *n* publishers in the market adopt e-publishing. The set of alternatives is then $A^{EE} = \{1, e1, 2, e2..., n, en\} \cup A_0$. All e-publishers determine two prices: A price for their traditional channel (p_i^{EE}) and a price for their Internet channel (p_{ei}^{EE}). Each e-publisher's aim is to define its optimum prices (p_i^{EE} and p_{ei}^{EE}) in the given market environment that maximizes its own profit:

$$\max_{p_i^{EE}, p_{ei}^{EE}} \left(\prod_i^{EE}\right) = \max_{p_i^{EE}, p_{ei}^{EE}} \left(p_i^{EE} - c_i\right)\left(M.\frac{e^{-b.p_i^{EE}}}{1+\sum_{t=1}^{n} e^{-b.p_t^{EE}} + e^{-\vartheta b.p_{ei}^{EE}}}\right)$$

$$+\left(p_{ei}^{EE} - c_{ei}\right)\left(M.\frac{e^{-\vartheta b.p_{ei}^{EE}}}{1+\sum_{t=1}^{k} e^{-\vartheta b.p_{ei}^{EE}} + e^{-b.p_i^{EE}}}\right)$$

$$\text{s.t. } p_i^{EE} - c_i \geq 0, \quad \forall i$$

$$p_{ei}^{EE} - c_{ei} \geq 0, \quad \forall i$$

$$p_i^{EE} \geq 0, \; p_{ei}^{EE} \geq 0, \quad \forall i \qquad (8)$$

The convexity of the maximization problem cannot be demonstrated, the sub-optimal solutions are computed using the GA. The unit costs of p-publishing and e-publishing are assumed to be the same for two e-publishers and they are set to 1 and 0.75, respectively. The equilibrium price values and related demands and profits are given in Table 3. Both e-publishing and p-publishing prices of *e-publisher*$_2$ are lower than the ones of *e-publisher*$_1$. The reason is that both types of readers (e-readers and p-readers) of *e-publisher*$_2$ are more sensitive to price than the readers of *e-publisher*$_1$. *E-publisher*$_2$ is obliged to hold its prices down in order to grab more readers.

TABLE III. PRICE, DEMAND AND PROFIT VALUES AT EQUILIBRIUM IN SCENARIO 3

	Equilibrium values			
	e-publisher$_1$		*e-publisher*$_2$	
	e-publishing	*p-publishing*	*e-publishing*	*p-publishing*
Offered price	1.3775	2.1349	1.3210	1.7446
Profit	13.3059		7.4673	
Demand	12.9251	6.9969	7.3918	7.5115
Total profit	20.7732			

Based on the numerical results of three demonstrative examples, it is possible to make some observations. The scenario where the sum of the profits of all the publishers in the market is maximum is the third one: the E-E competition. The E-E competition can be interpreted as the industry equilibrium. The results show that the coefficient of price sensitivity (*b*) is the fundamental factor that determines both types of price levels. As the price sensitivity of readers increases, publishers are obliged to decrease their prices in order to maintain more reader. In this sense, it is the reader that determines the price level. In the case *e-publisher*$_1$ is alone in the market (E-P competition), its demand and accordingly its profit are higher than in the competitive market (E-E competition). In the case there is at least one e-publisher in the market, the p-publisher loses profit. The p-publisher cannot reduce price like an e-publisher because of its higher maintenance and stocking costs. An e-publisher's profit reaches its maximum level, if its competitor is a p-publisher.

IV. CONCLUSIONS

With the proliferation of smart mobile phones, such as iPhone or Blackberry, or the dedicated e-readers, such as Kindle or Nook, more and more young readers are accustomed to reading the electronic versions of the books and magazines. This has an increasing effect on the adoption rate of e-publishing; however it creates a pressure on the publishers. Now, publishers need to analyze the market more deeply in order to maintain their customer base. The principle objective of this work is to propose a strategic analysis framework which would be valid for different types of decision environments by making slight modifications. The concentration is on the publishing industry, especially on a p-publisher that tries to decide on whether to enter the e-publishing business. However, the same decision support framework can help an existing firm when changing its business, or an existing firm to determine new prices, or a new entrant firm to determine prices. The framework is built on game theory basis, since a successful business strategy needs to consider the actions of other players in a competitive market. Using a local search algorithm, the GA, in order to find the equilibrium points of the games is the other contribution of this work. The numerical results have shown that GA can found similar results with a non-linear solver. The results have revealed that the price sensitivity of readers is the most fundamental parameter that affects both the choice of the reader and the price of the book. The higher the reader's price sensitivity is, the more the publisher tends to sell e-book to the reader.

Going forward, the equilibrium points may be found by using other soft computing algorithms and different algorithms can be compared in terms of obtained values and the execution time. Since the games are played offline in the proposed framework, the execution times are not questioned. The price sensitivity coefficient has a direct and profound impact on the equilibrium results, a future work can concentrate on determining it in the most efficient way. Some regression techniques can be useful in the presence of real life data.

ACKNOWLEDGMENT

This research has been financially supported by Galatasaray University Research Fund.

 ISBN: 978-1-61208-181-6

REFERENCES

[1] G. Hua, T.C.E. Cheng, S. Wang, "Electronic books: To "E" or not to "E"? A strategic analysis of distribution channel choices of publishers," International Journal of Production Economics 2011, 129, pp. 338-346.

[2] J.H. Holland, Adaptation in natural and artificial systems, Ann Arbour: University of Michigan Press, 1975.

[3] N.R. Siriwardene and B.J.C. Perera, "Selection of genetic operators for urban drainage model parameter optimization," Mathematical and Computer Modelling 2006, 44, pp. 415-429.

[4] T. Riechmann, "Genetic algorithm learning and evolutionary games," Journal of Economic Dynamics & Control 2011, 25, pp. 1019-1037.

[5] J. Periaux, H.Q. Chen, B. Mantel, M. Sefrioui and H.T. Sui, "Combining game theory and genetic algorithms with application to DDM-nozzle optimization problems," Finite Elements in Analysis and Design 2001, 37, pp. 417-429.

[6] S. Hamblin and P.L. Hurd, "Genetic algorithms and non-ESS solutions to game theory models," Animal Behaviour 2007, 74, pp. 1005-1018.

[7] Y. Jiang and E. Katsamakas, "Impact of e-book technology: Ownership and market asymmetries in digital transformation," Electronic Commerce Research and Applications 2010, 9, pp. 386-399.

[8] F. Bernstein, J.S. Song and X. Zheng, "Bricks-and mortar" vs. "clicks and mortar": An equilibrium analysis," European Journal of Operational Research 2008, 187, pp. 671-690.

[9] T. Başar and G.J. Olsder, Dynamic Noncooperative Game Theory. Academic Press, 2e, 1995.

[10] K.T. Talluri and G.J. Van Ryzin. The Theory and Practice of Revenue Management. Kluwer Academic Publishers, 2004, Boston.

[11] J. Nash, "Non-Cooperative Games," The Annals of Mathematics 1951, 54(2), pp. 286-295.

 ISBN: 978-1-61208-181-6

An Integrated Decision Support System for Selecting Software Systems

Tuncay Gürbüz[1], S. Emre Alptekin[1], Gülfem Işıklar Alptekin[2]

Galatasaray University
[1]Department of Industrial Engineering
[2]Department of Computer Engineering
Istanbul, Turkey
tgurbuz@gsu.edu.tr; ealptekin@gsu.edu.tr; gisiklar@gsu.edu.tr

***Abstract*- Enterprise Resource Planning (ERP) product selection can be seen as one of the most critical and difficult decision making stages for an organization. This research explores the application of a hybrid multi-criteria decision making (MCDM) procedure for the evaluation of various ERP alternatives. The proposed evaluation framework integrates three methodologies: Analytic Network Process (ANP), Choquet Integral (CI) and Measuring Attractiveness by a Categorical Based Evaluation Technique (MACBETH). ANP produces the priorities of alternatives with respect to the interdependent evaluation criteria. The conjunctive or disjunctive behaviors between criteria are determined using MACBETH and CI. Numerical application of the proposed methodology is implemented on the decision making problem of a firm that faces with four ERP projects. The final ranking is compared to the one obtained by ignoring the interactions among criteria. The results demonstrate that the ignorance of the interactions may lead to erroneous decisions.**

Keywords- ERP, supplier selection, MCDM, ANP, Choquet Integral.

I. INTRODUCTION

An ERP system is a critical investment that can significantly influence future competitiveness and performance of a firm. It is increasingly important in today's modern businesses because of its ability to integrate the flow of material, finance, and information and to support organizational strategies [1]. A successfully implemented ERP can offer organizations automating business process, timely access to management information and improving supply chain management through the use of e-commerce [2]. It standardizes processes and stores information as well as recalls that data when it is required in real time environment. Implementing an ERP system may be costly and time-consuming. Companies spend billions of dollars and use numerous amounts of man-hours for installing elaborate ERP software systems. However, the benefits of a successful ERP project are worthwhile. In order to implement an ERP software successfully, it is necessary to select an ERP system which can be aligned with the needs of the company. Thus, an efficient decision making approach for ERP software selection requires both company needs and characteristics of the ERP system and their interactions to be taken into account [3]. The selection process for determining the most appropriate ERP software among a set of possible alternatives in the market is a multi-criteria decision making problem.

This paper introduces a hybrid multi criteria decision making (MCDM) model for ERP selection based on Analytic Network Process (ANP), Choquet Integral (CI) and Measuring Attractiveness by a Categorical Based Evaluation Technique (MACBETH). Initially, we have categorized ERP selection criteria into three main criteria set: Vendor related criteria (VRC), customer related criteria (CRC), and software related criteria (SRC). Each one consists of its own sub-criteria set. Since these sub-criteria and criteria have both inner and outher dependencies, we have made use of the ANP to determine these dependencies and relative priorities of all criteria. MACBETH is both an approach and a set of techniques that have the goal of providing an overall ordering of options, and that aid on the construction of interval numerical scales based on qualitative (non-numerical) pairwise comparison judgments [4, 5]. In this research, we have used MACHBETH and CI to determine conjunctive or disjunctive behaviors between criteria. The last phase of the proposed methodology involves ranking the given ERP alternatives according to their final performance scores. We have shown the feasibility of the proposed framework on the decision making problem of a company that needs to evaluate four ERP software alternatives and select the most suitable one according to its requirements.

The remaining part of the paper is organized as follows: in Section 2 we give related literature. Section 3 briefly describes the methodologies that constitute the proposed framework. The steps and details of the proposed decision support framework and its implementation into the ERP selection problem is given in Section 4. Finally, Section 5 concludes the study.

II. LITERATURE REVIEW

There are various solutions in the ERP market and each one of them has its own features. Furthermore, buying an

 ISBN: 978-1-61208-181-6

ERP software is a very serious and difficult task for a firm since it may cost hundreds of thousands and even millions of dollars. Therefore, in the academic literature there are many research on the selection process of ERP products. Wei et al. [1] have presented a comprehensive framework for selecting a suitable ERP system. Their selection framework is based on analytic hierarchy process (AHP) approach. Liao et al. [2] have presented a model which is based on linguistic information processing, for dealing with such a problem. Yazgan et al [6] have considered this selection problem as a multi-criteria decision making problem and they have integrated artificial neural network and analytic network process. This integration enables them to interview only one expert for the assessments. Karsak and Özoğul [3] have proposed a selection framework that enables both company demands and ERP system characteristics to be considered, and provides the means for incorporating not only the relationships between company demands and ERP system characteristics but also the interactions between ERP system characteristics through adopting quality function deployment principles.

III. THE METHODS

A. Analytic Network Process (ANP)

ANP is a generalization of Saaty's AHP, which is one of the most widely used multi-criteria decision support tools. AHP is limited to relatively static and unidirectional interactions with little feedback among decision components and alternatives [7].

Many real life decision problems cannot be structured as a hierarchy because of the fact that they involve the interaction and dependence of higher level elements in a hierarchy on lower level elements. So the hierarchy becomes more like a network. On this context, ANP and its supermatrix technique can be considered as an extension of AHP that can handle a more complex decision structure as the ANP framework has the flexibility to consider more complex interrelationships (outerdependence) among different elements [8, 9].

AHP incorporates both qualitative and quantitative approaches to a decision problem [10]. It is also capable of capturing the tangible and intangible aspects of relative criteria that have some bearing on the decision making process, but AHP cannot deal with interconnections and innerdependences between decision factors in the same level [8]. This is because an AHP model is structured in a hierarchy in which no horizontal links are allowed. In other words, AHP can only be applied to a hierarchy that assumes unidirectional relation between decision levels. In fact, this weakness can be overcome by using the advance multi-criteria making technique, which is ANP. So, ANP is very useful in these kinds of situations providing a general framework without the assumptions of independence of higher-level elements from lower ones, or independence on the same level [11].

In this approach, comparison matrices, prioritization and the weights while considering the interdependencies are formed between various attributes of each level with the scale of 1–9 suggested by Saaty [12]. Also the consistencies of the pairwise comparisons, made by the experts or decision makers, have to be checked in order to make the necessary changes if there is any inconsistency above the allowed limit. Once the pairwise comparison matrices are formed, weighted vectors for all the matrices are calculated. The concept of supermatrix is employed to obtain the composite weights that overcome the existing interrelationships. The synthesizing step is to rate the alternatives according all the criteria, compute the overall score for the alternatives and make the final decision as to choose the best alternative or to obtain the final ranking of the alternatives.

B. Choquet Integral (CI)

The CI, which has been introduced in the fuzzy measure community by "Murofushi and Sugeno [13]" is a fuzzy integral proposed by "Gustave Choquet [14]" and considers the interactions between k out of n criteria of the problem, which is called the k-additivity property.

Letting $t_i, \quad i = 1, \ldots, n$ be the scores on the criteria, by using only the interaction index, it is possible to express CI in the case of 2-additive measures as follows [15]:

$$C_\mu(t_1, \ldots, t_n) = \sum_{I_{ij}>0} (t_i \wedge t_j) I_{ij} + \sum_{I_{ij}<0} (t_i \vee t_j) \left| I_{ij} \right| + \\ + \sum_{i=1}^{n} t_i (\Phi_i - \frac{1}{2} \sum_{j \neq i} \left| I_{ij} \right|) \qquad (1)$$

$$\text{with} \quad \Phi_i - \frac{1}{2} \sum_{j \neq i} \left| I_{ij} \right| \geq 0, \quad \forall i = 1, \ldots, n$$

Here, Φ_i represents the relative importance of criterion i with $\sum_{i=1}^{n} \Phi_i = 1$ and I_{ij}, defined in the interval [-1; 1], is the interaction value between criteria i and j. Positive values of I_{ij} implies a conjunctive behavior between criteria i and j. *i.e.* simultaneous satisfaction of both criteria is significant for the global score. Negative values of I_{ij} implies a disjunctive behavior between criteria i and j. *i.e.* the satisfaction of either one is sufficient to have a significant effect on the global score. If I_{ij} is null, then there is no interaction between criteria i and j. If for all pairs of criteria, I_{ij} are null then the Φ_i value acts as a weight vector in a weighted arithmetic mean.

C. Measuring Attractiveness by a Categorical Based Evaluation Technique (MACBETH)

MACBETH is a multi-criteria decision analysis approach which has been proposed by studies of "Bana e

Costa [16]", "Bana e Costa & Vansnick [17]", "Bana e Costa & Vansnick [18]". The method requires only qualitative judgments about differences of value to help an individual or a group in quantifying the relative attractiveness of the elements of a finite set *A* and to associate a real number *v(x)* to each element *x* of *A* [19].

Let *X* be the finite set of elements (alternatives) with at least two elements and *J* the group of decision makers who want to compare the relative attractiveness of these elements. Here, it is assumed that the DM or each DM is able to rank the elements of *X* either directly or through pairwise comparisons. Each DM is first asked to provide a judgment about the relative attractiveness of two elements at a time to retrieve the ordinal judgment. Then secondly, he/she is asked to provide a qualitative judgment on the difference of attractiveness of those two elements if they are not equally attractive using the following linguistic terms: *Very weak*, *weak*, *moderate*, *strong*, *very strong* and *extreme*.

MACBETH method presents a procedure to transform qualitative preferences into coherent quantified elementary and aggregated performances. In order to solve the inter-criteria commensurability problem, it is sufficient to determine, for all interval scales, two common reference points namely the good situation and the neutral situation with the performance values one and zero respectively.

Let p_i^k be the performance expression of the k^{th} alternative for criterion *i*. Suppose the DM prefers for criterion *i* the alternative *k* to the alternative *l* and in addition to that information, DM will characterize the strength of his judgments with a level of strength that can take values from one to six (from the least to the most strong level) according to the six semantic categories of difference of attractiveness explained above and zero for a null strength. This level will be denoted with *h*. Therefore, if the DM prefers for criterion *i* the alternative *k* to the alternative *l*, with a strength *h*, then the following equation, where α is a coefficient necessary to meet the condition p^k and $p^l \in [0;1]$, will be obtained:

$$A^k \succ^h A^l \Leftrightarrow p_i^k - p_i^l = h\alpha \quad (2)$$

Therefore, a preference ranking of alternatives for a specific criterion collected from a DM with the strength of the comparisons will give us a system of equation and after solving it the individual performance values of the alternatives for the criterion in question will be determined.
In order to define CI parameters, the DM is asked to provide preferential information on the criteria and the couples of criteria including the strength of the preferences. This information will help us to build a system of equations with the Shapley and the Interaction parameters as variables. For the elementary performance expressions, MACBETH proposes to consider some particular and possibly fictive situations, *S*, in which the alternatives satisfies one criterion or two criteria simultaneously. A preference ranking of those situations collected from a DM with the strength of the comparisons will give us a system of equation and after solving it the CI parameters will be determined.

In the situations where only one $p_i = 1$ (i.e. criterion *i* is satisfied) and all others are equal to zero, the aggregated performance expression will be as follows (Cliville et al. 2007):

$$p_{Ag}^i = \varphi_i - \frac{1}{2}\sum_{\substack{j=1 \\ j\neq i}}^{n} I_{ij} \quad (3)$$

The aggregated performance expression of the situations where only one $p_i = 0$ and all others are equal to one (i.e. all criteria except *i* is satisfied) will be as follows [5]:

$$p_{Ag}^i = 1 - \varphi_i - \frac{1}{2}\sum_{\substack{j=1 \\ j\neq i}}^{n} I_{ij} \quad (4)$$

The aggregated performance expression of the situations where only two elementary performance expressions are equal to one (namely *i* and *j*) and all others are equal to zero (i.e. criterion *i* and *j* are satisfied) will be as follows [20]:

$$p_{Ag}^{i,j} = \varphi_i + \varphi_j - \frac{1}{2}\left(\sum_{k\in\aleph_{1,n}\,|\,p_k=0} I_{ik} + \sum_{k\in\aleph_{1,n}\,|\,p_k=0} I_{jk} \right) \quad (5)$$

IV. ERP SYSTEM EVALUATION FRAMEWORK

A. Evaluation procedure

The evaluation procedure of this study consists of seven steps as follows:

1. Identify the ERP software selection/ evaluation criteria that are considered the most important for the users.
2. Once the model is built and the relations between criteria are defined, decide the method to use. This is not an arbitrary choice.
3. If there is an outer-dependence between sub-criteria, then this is something to be analyzed with ANP because of the simple fact that CI cannot handle two elements that are connected to two different points. In this case, two sub-criteria in question belong to two different criteria. Hence, these dependencies will be handled with ANP.
4. Analyze sub-criteria of the same cluster in order to define the conjunctive and disjunctive behavior between them. If there is such relation, use CI in order to find the interaction values. In case of no such interaction, handle the relations with ANP.
5. After handling the sub-criteria, take in consideration the upper level, i.e. the criteria.
6. A preference ranking of the criteria given by the DMs will define the conjunctive/disjunctive behavior between those. If there is not any interaction of this kind between criteria, then solve the model with ANP. Make the final aggregation and obtain a ranking.
7. If there are conjunctive/disjunctive behavior between criteria, then use the Shapley indices and the interaction

 ISBN: 978-1-61208-181-6

values including the weights of the sub-criteria and the alternatives' individual performance values for each of those sub-criteria in order to perform the final aggregation.

B. *Selection criteria*

Baki and Çakar have summarized the ERP selection criteria in their research after reviewing the related literature [21]. We have used the 16 criteria that they have proposed but we have grouped them under three main categories: vendor related, customer related and software related (Table 1).

TABLE I. SELECTION CRITERIA

C_1	**Vendor related criteria (VRC)**
C_{11}	Support and service
C_{12}	Vision
C_{13}	Market position
C_{14}	Domain knowledge
C_{15}	Reputation
C_{16}	Methodology of software
C_2	**Customer related criteria (CRC)**
C_{21}	Ease of customization
C_{22}	Better fit with organizational structure
C_{23}	Fit with parent/allied organizational system
C_{24}	Cross module integration
C_3	**Software related criteria (SRC)**
C_{31}	Functionality
C_{32}	Technical aspects
C_{33}	Cost
C_{34}	System reliability
C_{35}	Compatibility
C_{36}	Implementation time

C. *Proposed decision framework*

The hierarchical structure of the decision model of the paper with the alternatives and the identified criteria is portrayed in Fig. 4. The proposed decision model consists of three levels: at the highest level the objective of the problem is situated while in the second level, the criteria are listed. The lowest level belongs to the alternatives. As alternatives, A_1, A_2, A_3 and A_4 are selected since they are in the same interval of price.

D. *NUMERICAL APPLICATION OF THE PROPOSED FRAMEWORK*

1) *Part 1: ANP*

In the first part of the framework, pairwise comparison matrices for all the sub-criteria have been prepared and filled out by the DM. The consistency indexes of the matrices are all smaller than 0.10, which proves their consistency [12]. The pairwise comparisons enable us to retrieve relative weights for the sub-criteria. The supermatrix, which has the role of obtaining the composite weights, has been constructed (Table 2).

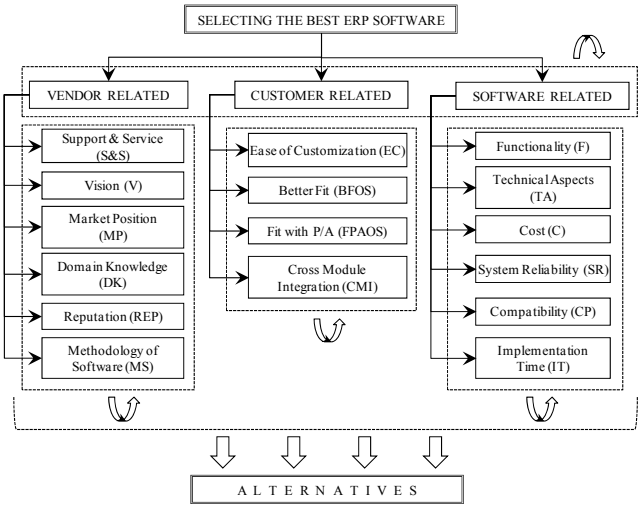


Fig. 4. Hierarchical structure of the decision making problem

TABLE II. UNWEIGHTED SUPERMATRIX.

	S&S	V	MP	DK	REP	MS	EC	BFOS	FPAOS	CMI	F	TA	C	SR	CP	IT
S&S	0	0	0.07	0	0.25	0	0.39	0.25	1	0	0	0	0.11	0	0	0.11
V	0.75	0	0.15	1	0.75	1	0.07	0	0	0	0.25	0	0	0.17	0.13	0
MP	0	0	0	0	0	0	0	0	0	0	0	0	0.64	0	0	0
DK	0.25	0	0.39	0	0	0	0.39	0.75	0	0	0.75	0	0	0.83	0.88	0.26
REP	0	0	0.39	0	0	0	0	0	0	0	0	0	0.26	0	0	0
MS	0	0	0	0	0	0	0.15	0	0	0	0	0	0	0	0	0.64
EC	0.64	0	0.43	0	0	0	0	0.43	0.25	1	0.08	0.19	0	0	0.15	0.38
BFOS	0.11	0	0.14	0	0	0	0.09	0	0	0	0.20	0.08	0	0	0.39	0.13
FPAOS	0	0	0	0	0	0	0.46	0.43	0	0	0.20	0.08	0	0.75	0.39	0.38
CMI	0.26	0	0.43	0	0	0	0.46	0.14	0.75	0	0.52	0.66	0	0.25	0.07	0.13
F	0	0	0.05	0	0	0	0.25	0	0	0	0	0	0.19	0.25	0	0.26
TA	0	0	0	0	0	0	0.75	0	0	1	0	0	0.07	0.75	1	0.11
C	0	0	0.48	0	0	0	0	0	0	0	0.73	0.64	0	0	0	0
SR	1	0	0.21	0	1	0	0	0	0	0	0	0.26	0.19	0	0	0
CP	0	0	0.05	0	0	0	0	0	1	0	0.18	0.11	0.07	0	0	0.64
IT	0	0	0.21	0	0	0	0	0	0	0	0.08	0	0.47	0	0	0

As the next step, cluster/criteria comparison matrices have been prepared and filled out by the DM in order to normalize the supermatrix (Table 2). Using the weights retrieved from these matrices, the cluster matrix is constructed and weighted supermatrix is calculated. The cluster matrix and the weighted supermatrix are represented in Table 3 and Table 4, respectively.

TABLE III. CLUSTER MATRIX.

	VRC	CRC	SRC
VRC	0.637	0.258	0.258
CRC	0.105	0.637	0.105
SRC	0.258	0.105	0.637

 ISBN: 978-1-61208-181-6

TABLE IV. WEIGHTED SUPERMATRIX.

	S&S	V	MP	DK	REP	MS	EC	BFOS	FPAOS	CMI	F	TA	C	SR	CP	IT
S&S	0	0	0.043	0	0.178	0	0.101	0.072	0.258	0	0	0	0.030	0	0	0.027
V	0.478	0	0.097	1	0.534	1	0.018	0	0	0	0.07	0	0	0.043	0.033	0
MP	0	0	0	0	0	0	0	0	0	0	0	0	0.184	0	0	0
DK	0.159	0	0.248	0	0	0	0.101	0.216	0	0	0.19	0	0	0.215	0.226	0.067
REP	0	0	0.248	0	0	0	0	0	0	0	0	0	0.075	0	0	0
MS	0	0	0	0	0	0	0.039	0	0	0	0	0	0	0	0	0.165
EC	0.067	0	0.045	0	0	0	0	0.305	0.159	0.87	0.01	0.027	0	0	0.016	0.039
BFOS	0.011	0	0.015	0	0	0	0.058	0	0	0	0.02	0.011	0	0	0.041	0.013
FPAOS	0	0	0	0	0	0	0.290	0.305	0	0	0.02	0.011	0	0.079	0.041	0.039
CMI	0.027	0	0.045	0	0	0	0.290	0.102	0.478	0	0.05	0.093	0	0.026	0.007	0.013
F	0	0	0.013	0	0	0	0.026	0	0	0	0	0	0.138	0.159	0	0.165
TA	0	0	0	0	0	0	0.079	0	0	0.14	0	0	0.052	0.478	0.637	0.067
C	0	0	0.124	0	0	0	0	0	0	0	0.47	0.547	0	0	0	0
SR	0.258	0	0.055	0	0.289	0	0	0	0	0	0	0.222	0.138	0	0	0
CP	0	0	0.013	0	0	0	0	0	0.105	0	0.12	0.090	0.052	0	0	0.406
IT	0	0	0.055	0	0	0	0	0	0	0	0.05	0	0.332	0	0	0

As the final step of ANP procedure, from the weighted supermatrix given in Table 4, the limit supermatrix has been calculated. The relative importances of the sub-criteria have been collected and then normalized with respect to the criteria (Table 5).

TABLE V. WEIGHTED SUPERMATRIX.

VRC	**S&S**	**V**	**MP**	**DK**	**REP**	**MS**
w	0.138	0.473	0.065	0.235	0.045	0.044
CRC	**EC**	**BFOS**	**FPAOS**	**CMI**		
w	0.422	0.049	0.202	0.327		
SRC	**F**	**TA**	**C**	**SR**	**CP**	**IT**
w	0.099	0.276	0.238	0.159	0.126	0.102

2) Part 2: CI

In order to find out criteria weights and conjunctive/ disjunctive interactions between criteria, a preferential ranking, including the strength of the judgment, has been asked to the DM. The following ranking has been obtained:

$$C_1 \,\&\, C_3 \succ^{M} C_1 \,\&\, C_2 \succ^{W} C_2 \,\&\, C_3 \succ^{M} C_1 \succ^{W} C_3 \succ^{S} C_2 \succ^{S} \text{"0"}$$

Using Eq. 2-5, the resolution of this equation system has given the results shown in Table 6.

TABLE VI. CHOQUET INTEGRAL PARAMETERS.

Parameter	Φ_1	Φ_2	Φ_3	I_{12}	I_{13}	I_{23}
Value	0.4375	0.2083	0.3542	0.0417	0	0.0417

Table 6 shows that Vendor Related Criteria is the most important cluster for the DM with a relative importance (Φ_1) of 0.4375, whereas Customer Related Criteria is the least important with a relative importance (Φ_2) of 0.2083. Furthermore, it is possible to state that an employee must be successful on both VRC and CRC in order to be considered successful for the DM. The same situation is valid for CRC and SRC, as the interaction values (I_{12} and I_{23}) are positive for those couples of criterion. On the other hand, there is no conjunctive/ disjunctive interaction between VRC and SRC as the interaction value for these (I_{13}) is null.

3) Part 3: Relative performance scores of alternatives

For each sub-criterion, the DM has been asked to give us a preferential ranking, including the strength of the comparisons. For example, forVRC, the following ranking has been obtained.

$$C_{11} \rightarrow Good \succ^{VW} A_4 \succ^{M} A_1 \succ^{M} A_2 \succ^{M} A_3 \succ^{M} Neutral$$

$$C_{12} \rightarrow Good \succ^{VW} A_1 \succ^{W} A_2 \succ^{M} A_3 \succ^{S} A_4 \succ^{S} Neutral$$

$$C_{13} \rightarrow Good \succ^{VW} A_4 \succ^{W} A_1 \succ^{M} A_2 \succ^{S} A_3 \succ^{S} Neutral$$

$$C_{14} \rightarrow Good \succ^{VW} A_1 \succ^{W} A_2 \succ^{M} A_4 \succ^{M} A_3 \succ^{S} Neutral$$

$$C_{15} \rightarrow Good \succ^{VW} A_1 \succ^{W} A_2 \succ^{M} A_4 \succ^{S} A_3 \succ^{S} Neutral$$

$$C_{16} \rightarrow Good \succ^{VW} A_1 \succ^{W} A_2 \succ^{M} A_3 \succ^{S} A_4 \succ^{S} Neutral$$

Similarly, rankings for CRC and SRC has been obtained. The equations systems have been retrieved from these three preference rankings, using Eq. 2. The resolution of these equation systems are summarized in Table 7.

TABLE VII. PERFORMANCE VALUES OF ALTERNATIVES WITH RESPECT TO SUB-CRITERIA.

	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	C_{21}	C_{22}	C_{23}	C_{24}	C_{31}	C_{32}	C_{33}	C_{34}	C_{35}	C_{36}
A_1	0.692	0.929	0.786	0.923	0.929	0.929	0.214	0.625	0.625	0.786	0.714	0.857	0.067	0.867	0.800	0.133
A_2	0.462	0.786	0.571	0.769	0.786	0.786	0.429	0.813	0.813	0.643	0.929	0.714	0.267	0.733	0.667	0.267
A_3	0.231	0.571	0.286	0.308	0.286	0.571	0.643	0.563	0.563	0.500	0.500	0.500	0.533	0.467	0.400	0.533
A_4	0.923	0.286	0.929	0.538	0.571	0.286	0.857	0.188	0.188	0.214	0.286	0.286	0.800	0.267	0.200	0.800

4) Part 4: Final aggregation

As the last step of the proposed procedure, the scores of the alternatives for each criterion have been calculated (Table 8).

TABLE VIII. PERFORMANCE VALUES OF ALTERNATIVES WITH RESPECT TO CRITERIA.

	VRC	**CRC**	**SRC**
A_1	0.8854	0.5042	0.5754
A_2	0.7232	0.5950	0.5803
A_3	0.4310	0.5760	0.4934
A_4	0.4877	0.4788	0.4467

In order to determine the final performance scores of four alternatives (Table 9), we have used the values in Table 6 and Table 8 and Eq. 2.

 ISBN: 978-1-61208-181-6

TABLE IX. FINAL SCORES OF ALTERNATIVES.

	A_1	A_2	A_3	A_4
Score	0.6867	0.6429	0.4786	0.4705

The result indicates that the final performance score of alternative A_1 is the highest (0.6867) and that of alternative A_4 is the lowest (0.4705). The fact that VRC has the greatest relative importance (Φ_1=0.4375) has played an important role for A_1 and A_2 to be ranked first two in the final ranking and for A_3 and A_4 to be ranked last two. Although A_2 has greater performance values for CRC and SRC, VRC was the defining crieteria for A_1 to be ranked first. The same situation is present between A_3 and A_4: A_4 has lower performance values with respect to CRC and SRC and greater performance value with respect to VRC. However in this case, A_3 is ranked before A_4. The reason is the fact that the differences between performance values with respect to CRC and SRC for those two alternatives are greater than that for A_1 and A_2.

V. CONCLUSION

Enterprise resource planning is a software application package that integrates internal and external management information across an entire organization. As there are various ERP software product in the market, a client needs to choose the product that uses less resources and produces more output. In other words, the client needs to choose the most appropriate product for its own requirements.

Our work presents a comprehensive framework for selecting a suitable ERP system based on an hybrid multi-criteria decision analysis process. The procedure consists of three methodologies: Analytic Network Process (ANP), Choquet Integral (CI) and Measuring Attractiveness by a Categorical Based Evaluation Technique (MACBETH). We have illustrated the applicability of the framework through a case study of the ERP software selection of a company. The proposed decision making framework is flexible enough to fit other sectors with some specific characteristics changes and to incorporate different criteria in the evaluation process.

ACKNOWLEDGMENT

This research has been financially supported by Galatasaray University Research Fund.

REFERENCES

[1] Wei, C.C, Chien, C.F. and Wang, M.J.J., 2005, "An AHP-based approach to ERP system selection", International Journal of Production Economics, 96, pp. 47-62.

[2] X. Liao, Y. Li and B. Lu, "A model for selecting an ERP system based on linguistic information processing", Information Systems, vol. 32, 2007, pp. 1005-1017.

[3] Karsak, E.E. and Ozogul, C.O. 2009, "An integrated decision making approach for ERP system selection", Expert Systems with Applications, 36, pp. 660-667.

[4] Bana e Costa, C.A., Chagas, M. P., 2004, A career choice problem: An example of how to use MACBETH to build a quantitative value model based on qualitative value judgments, European Journal of Operational Research, 153, 323 – 331.

[5] Clivillé, V., Berrah, L. and Mauris, G. 2007, "Quantitative expression and aggregation of performance measurements based on MACHBETH multicriteria method", International Journal of Production Economics, 105(1), pp. 171-189.

[6] Yazgan, H.R., Boran, S. and Göztepe, K., 2009, "An ERP software selection process with using artificial neural network based on analytic network process approach", Expert Systems with Applications, 36, pp. 9214-9222.

[7] Meade, L., & Sarkis, J., 1998. Strategy analysis of logistics and supply chain management systems using the analytical network process. Transportation Research E: The Logistics and Transportation Review, 34(3), 51–65.

[8] Saaty, T.L., Decision Making with Dependence and Feedback: The Analytic Network Process, RWS Publications, Pittsburgh, 1996.

[9] Saaty, R. W., 2003. Decision making in complex environments. Pittsburgh:Creative Decisions Foundation.

[10] Cheng, E.W.L, Li, H., Yu, L., 2004, The Analytic Network Process (ANP) Approach to Location Selection: A Shopping Mall Illustration, Construction Innovation, 5, pp. 83 – 97.

[11] Gencer, C., Gürpınar, D., 2007, Analytic Network Process in Supplier Selection: A Case Study in an Electronic Firm, Applied Mathematical Modeling, 31, pp. 2475 – 2486.

[12] Saaty, T.L., The Analytic Hierarchy Process, McGraw-Hill, New York, 1980.

[13] Murofushi, T., Sugeno, M., 1989, "An interpretation of fuzzy measure and the Choquet integral as an integral with respect to a fuzzy measure," Fuzzy Sets Systems, vol. 29, pp. 201–227, 1989.

[14] Choquet, G., 1953, "Theory of capacities," Annales de l'Institut Fourier, vol.5, pp. 131–295.

[15] Grabisch, M., 1997, k-order additive fuzzy measures, 6th International Conference on Information Progressing and Management of Uncertainty, Fuzziness and Knowledge Based Systems, 5:587 – 607.

[16] Bana e Costa, C.A., 1992, Structuration, Construction et Exploitation d'un Modèle à Multi-critère d'aide à la Décision, PhD Thesis, Technical University of Lisbon, Lisbon,1992.

[17] Bana e Costa, C.A., Vansnick, J.C., 1993, Sur la Quantification des Jugements de Valeur: L'Approche MACBETH, Cahiers du LAMSADE, 117, Université Paris – Dauphine, Paris, 1993.

[18] Bana e Costa, C.A., Vansnick, J.C., 1994, MACBETH – An Interactive Path Towards the Construction of Cardinal Value Functions, International Transactions in Operational Research, Vol.1(4), pp.489 – 500.

[19] Bana e Costa, C.A., De Corte, J.M., Vansnick, J.C., 2005. On the mathematical foundations of Macbeth. In: Figueira, J., Greco, S., Ehrgott, M. (Eds.), Multiple Criteria Decision Analysis: State Of The Art Surveys. Kluwer Academic Publishers, Dordaecht, pp. 409–442.

[20] Gürbüz, T., (2010), Multiple Criteria Human Performance Evaluation Using Choquet Integral, International Journal of Computational Intelligence Systems, 3(3), 290 – 300.

[21] Baki, B. and Çakar, K. "Determining the ERP package-selecting criteria, The case of Turkish manufacturing companies", Business Process Management Journal, vol. 11(1), 2005, pp. 75-86.

 ISBN: 978-1-61208-181-6

Virtual World Process Perspective Visualization

Ross Brown
Queensland University of Technology,
Faculty of Science and Technology
Brisbane, Australia
e-mail: r.brown@qut.edu.au

Johannes Herter, Daniel Eichhorn
Institute for Information Management in Engineering,
Institute for Applied Informatics and Formal
Description Methods,
Karlsruhe Institute of Technology,
Karlsruhe, Germany
e-mail:{johannes.herter},{daniel.eichhorn}@kit.edu

***Abstract*— Product Lifecycle Management has been developed as an approach to providing timely engineering information. However, the number of domain specializations within manufacturing makes such information communication disjointed, inefficient and error-prone. In this paper we propose an immersive 3D visualization of linked domain-specific information views for improving and accelerating communication processes in Product Lifecycle Management. With a common and yet understandable visualization of several domain views, interconnections and dependencies become obvious. The conceptual framework presented here links domain-specific information extracts from Product Lifecycle Management systems with each other and displays them via an integrated 3D representation scheme. We expect that this visualization framework should support holistic tactical decision making processes between domain-experts in operational and tactical manufacturing scenarios.**

Keywords-Product Lifecycle Management; Process Modeling; Scientific Visualization

I. INTRODUCTION

The ability to deal with the increasing complexity of product manufacturing depends on mutual understanding and close collaboration between all manufacturing related domains. This level of collaboration is necessary in present day manufacturing enterprises, in order to make optimal decisions about product construction and process scheduling. This collaboration requirement leads to a strong demand for suitable information management and presentation tools as a fundamental component of successful operational and tactical decision making processes. In order to accelerate and improve holistic, integrated decision making, a conceptual framework is proposed, to provide multi-domain information visualization in a product engineering environment. This paper is structured as follows: First of all, the motivation, background and related work are introduced. On this basis, the conceptual framework is described by discussing several visualization options. Thereby special consideration is given to visualization approaches, which display links between central elements of information sets. This is concluded by a detailed analysis of hierarchical product structures and their impact on visualization.

A. Motivation and Problem Description

In recent years, Product Lifecycle Management (PLM) solutions have been developed for providing engineering related information wherever it is required. In PLM Systems, as an extension of Product Data Management (PDM), all relevant information relating products, resources and processes are stored [1], [2]. The internal data structure for storing and managing this information is referred to as the "Global Product Structure" (GPS) [3] in the context of this work. As the GPS contains all technical information of several engineering-related disciplines for all products and variants, the number of elements in the structure can easily become very large. In order to have a subset of information which is understandable and manageable, domain-specific views are created for the demands of specified target audiences. A view in this context is an extract from a global information system for domain experts containing a subset of relevant information in a suitable notation [2]. These subsets, extracted from the GPS, are the information used for decision processes and require inter-domain communication for making holistically optimized decisions. The conceptual framework introduced here is based on an elementary use case for a manufacturing decision, introduced in the following. The simplicity of the use-case regards the conceptual character of this approach; more complex scenarios are considered in future research.

In this scenario, a decision making process is presented concerning the domains of resource planning, manufacturing process planning and product construction. Between representatives of these domains, a commitment should be made to have an overall optimized (pareto-optimal) outcome. This use-case describes the discussion about an attachment process in order to manufacture a wooden bench. One manufacturing alternative would be gluing the attachment; the other is screwing the attachment. Each variant would bring benefits and disadvantages for the considered domains: construction, process scheduling and resource planning.

For benchmarking both process variants, performance indicators like manufacturing time, physical robustness, reusability or demand for several kinds of resources are identified. Following these indicators, it is plausible that the optimal choice of method could be different for each domain. In order to achieve an overall optimal choice, the

 ISBN: 978-1-61208-181-6

benefits of each variant have to be negotiated from the point of view of all disciplines in a process of direct communication. This discussion would end in choosing one of the variants, which would bring more benefits than disadvantages when considering all regarded disciplines.

II. Background Theory and Previous Work

Following "General Model Theory" by Stachoviak [4] the models for each discipline, created with the motivation of "Pragmatism", contain elements which are only relevant in the corresponding domain. The individual models form the basis for domain representations in decision making processes. The fundamental issue for these cross domain negotiations is the development of domain-specific models for each participant. The direct communication between representatives of participating domains should be supported by a suitable visualization, which integrates and shows information extracted from all domains. Thereby, links and dependencies between all related models become obvious, communication processes are accelerated and consequently, decision quality is improved.

Here the dilemma becomes obvious: within a common visualization on classic displays, the understandability of the extracts get lost as the number of items and their interconnections can easily get too high when the illustration is composed of several domain-specific extracts on one flat screen. Furthermore, there might be syntactic and semantic differences between the domain-specific models which could hinder understanding a combined visualization.

To overcome the dilemma of complexity and understandability, a 3D visualization is proposed in this paper, to display the information from various domains simultaneously in an understandable and intuitive manner.

In the field of visualization of abstract data sets, related work focuses information from singular domains like, e.g., process management [6], [9], [13], [16], [17]. In the field of visual data-mining, 3D visualization is used for showing unknown correlations between data [13]. Other approaches conduct 3D visualization for large data sets [18]. However, a 3D visualization for simultaneously showing information models from different domains seems not to be in the focus of research.

Human spatial visualization capabilities give the opportunity to place domain-specific views in a 3D virtual space. Users can see domain-specific information models in the foreground or background, according to their position and orientation in the virtual space. This fact alone motivates the development of visualizations within a Virtual Reality (VR) environment, to increase understandability by full immersion and interaction, and has indeed led to an initial implementation. An initial VR prototype is under development in the Lifecycle Solutions Center (LESC), in Karlsruhe Institute of Technology, Karlsruhe, Germany.

A process-oriented use-case has been chosen to illustrate this integrated visualization challenge. The manufacturing process is used as the major information model with which other domain-specific data-structures are linked. A process oriented approach, from industry and science, is used, due to its relevance in product manufacturing [5].

III. 3D Domain Visualization Approach

In order to integrate and show cross domain PLM data in a coherent manner, we propose to use a 3D Cylinder structure modified to suit the particular requirements of PLM process data [6], see Fig. 2.

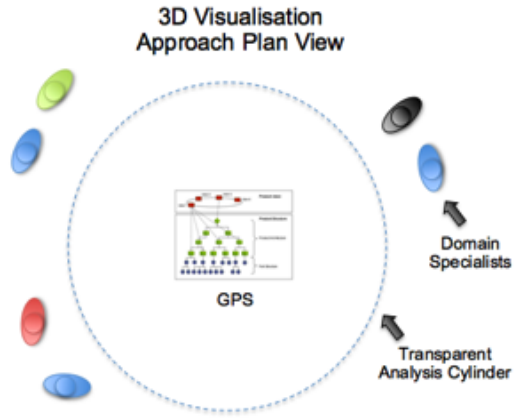


Figure 1: Plan View of 3D Cylinder-based PLM visualization structure with the outer cylinder used for analysis, and the inner GPS structure to provide context with regards to the entire GPS.

A plan view of the cylinder-based representation is shown in Fig. 1, using an inner representation of the GPS data structures related to a production process and an outer cylinder surface devoted to showing analytical representations. The intention is that such a representation is used within a fully immersive VR facility, as previously mentioned. Using this visualization approach, production domain specialists are able to stand around the cylinder and view domain relevant analytic representations, seeing how they relate to the inner data structure used as a representation of the PLM processes. This approach provides two key functions, *Context* and *Collaboration*.

Context is provided by a representation of the PLM GPS within the cylinder, showing the GPS in entirety, using appropriate representations, supports traversal of its large hierarchical data structures. Or conversely, the interior GPS representation may be via a visualization of the product itself. This internal representation will be described in detail later.

Secondly, the analytic surface facilitates collaboration and comparison via the close juxtaposition of multiple, different, domain representations. Domain specialists may gather around and compare values, based upon the domain representations that suit each viewer, enabling cross-domain discussion and collaboration.

We consider that a full virtual reality representation of the cylinder would work best, due to the scope of the visualization lending itself to the traversal of large scale data sets, such as those found in manufacturing [2], [3]. The

 ISBN: 978-1-61208-181-6

visualizations in this paper are mockups shown in a virtual world that indicate the scale of the system to be devised by using an avatar for size comparison. It also should be noted that the representations on the surfaces of the cylinder in the figures are images used to convey the basis of the future implementation, and so should be considered proof of concept in nature.

Interactions with the cylinder system occur at two main points, with the external structure, via rotating the outer layer around the *y* axis, and by selecting and filtering information from the internal GPS structure. One person in the analysis team would typically be in control of these interactions and visual representations used in the visualization.

The outer layer can be spun, and is intended to be related to the internal GPS via visible linkages, shown as red arrows in our example figures. The internal GPS is interactive and can be used as an information source, to extract a domain-specific representation to be displayed on the outer cylinder. We now discuss the representations in detail.

A. Cylinder Surface Related Representations

For domain analysis purposes, different representations are applied from the central GPS from a PLM System onto an outer cylinder. We add domains and processes into the top levels above the views, with indicators of differences in the graphs [7] for comparisons of processes and their influences on the results of the PLM processes. The flow on cost effects can be viewed in the lower level, or over the elements of the process model, applied to another part of the cylinder surface.

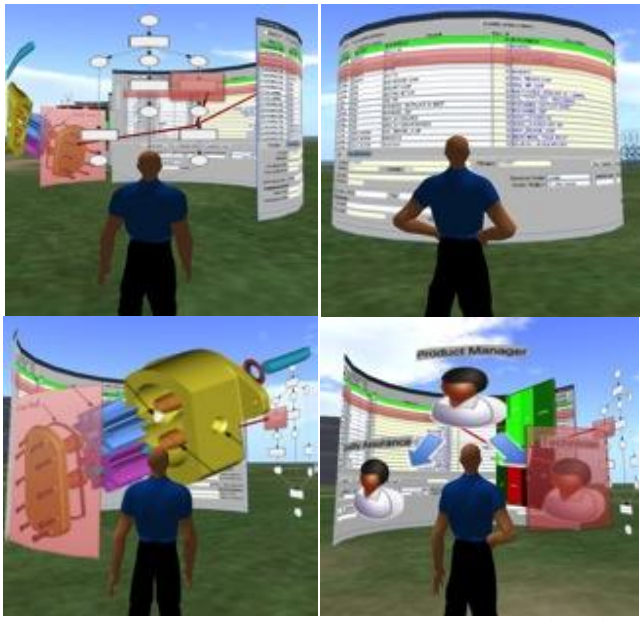


Figure 2: Analysis representations that can be applied to the surface of the cylinder regarding: Production Processes (top left), Bill Of Materials (BOM) (top right), Product Schematics (bottom left) and Human Resources (bottom right). Red highlighted areas indicate links to select information in the central GPS.

Each of these can be either drilled down upon in the central GPS representation, or specifically selected by the viewing session controller using a search interface. Models describing a certain region of interest are extracted from the GPS, following the pragmatism (demand for information) of several domains. Interconnections are taken from the PLM System, where the links are stored within the Global Product Structure [3], augmented with process linkages, shown in Fig. 3.

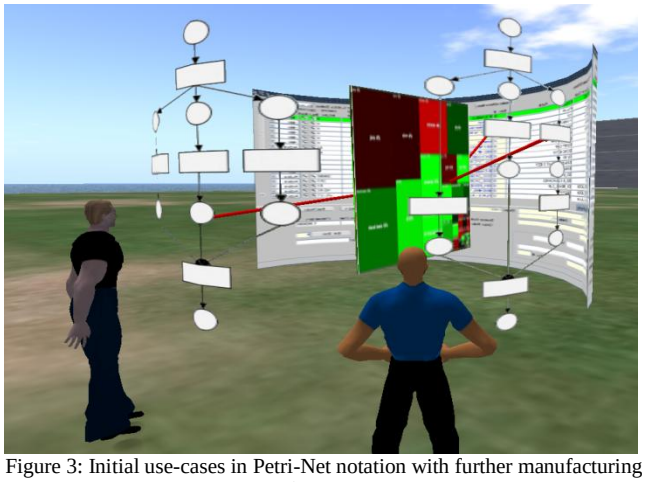


Figure 3: Initial use-cases in Petri-Net notation with further manufacturing information.

The extracted models are then processed to identify grouping relationships or pivotal elements. These elements are taken as links between the information models. Elements of the models are linked with 3D representations following defined mapping specifications between the abstract model and graphical representations. The mapping specifications include a rule-set for spatial arrangements, to assist in providing an optimal layout.

B. Internal Cylinder GPS Data Representations

We propose two internal representations for the PLM data: Tree-Map Hierarchy Stack and a Product Centric view. Each of these is chosen to represent the complex production data in a manner that provides context and complexity reduction when seeking to traverse and extract other representations to be displayed on the outer surface of the cylinder for domain analysis. Each of these representations also provides the key abilities of context, drill down and interaction required to traverse wide and deep GPS hierarchy information.

1) Tree-Map Hierarchy Stack: The Tree-Map Hierarchy Stack representation consists of layers from each level of the GPS, summarized to provide context data. Such an overview is often used for strategic views of data in other applications enabling parallax at each level to easily identify the planar position of objects. Such an approach provides complexity management by the spatial arrangement of the objects in a manner so that constrained camera motion reinforces the categories instinctively. This approach has been applied to large-scale process visualization approaches in Biology and we expect this approach to scale to large PLM Systems with multiple sources of data.

 ISBN: 978-1-61208-181-6

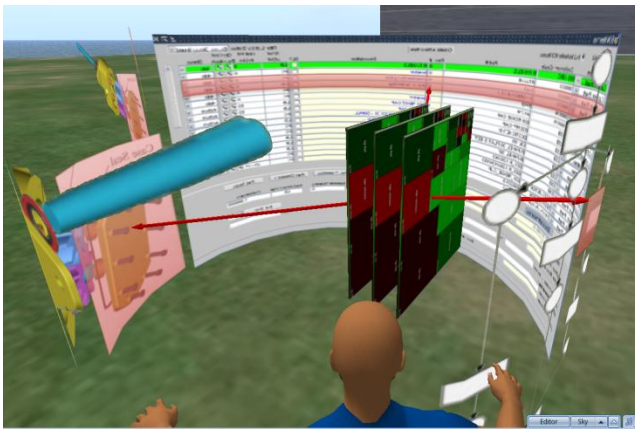


Figure 4: Example showing the inner structure as a Hierarchy Stack, made up of summarized Tree-Map levels within the GPS.

Each level of the hierarchy is composed of a Tree-Map structure. The Tree-Map GPS representation (refer to Fig. 5) is a popular 2D embedded hierarchy representation that can facilitate display and traversal of large hierarchical data structures, showing the path through the hierarchy, and allowing drill down capabilities [8]. This representation will allow domain specialists to see the contents of the GPS levels at a glance, and still enable interactive sense making of the relationships between each component. In particular, the path through the hierarchy can be traced across the surface of the Tree-Map, so its contextual capabilities give it an ability to facilitate transitions to areas of interest (see Fig. 7).

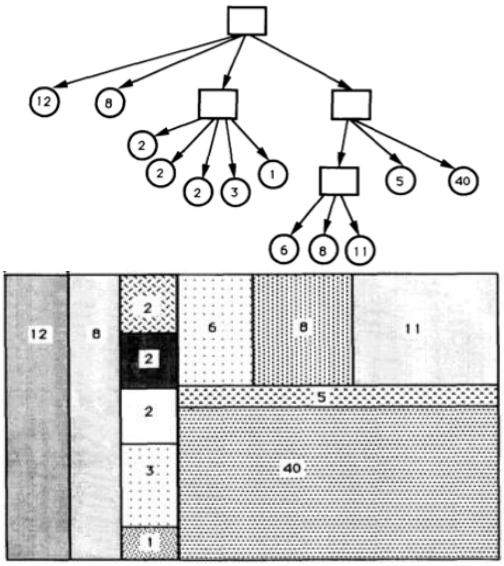


Figure 5: Simple example of a Tree-Map converting a tree structure (top) represented as a map structure (below) from [8].

2) Product Centric: A product centric representation (refer to Fig. 6), provides a hierarchical representation of the production process, from the point of view of the physical product components. 3D exploded representations of the product are shown in the central area of the cylinder, as a form of hierarchical representation. This representation is expected to be useful as a domain specialist independent view of the GPS, as it can be reasonably assumed that all stakeholders in a production facility are able to identify physical products and their components.

Figure 6: Example using a product representation as the inner structure.

The GPS views can be constructed by attribute filtering, architecture references, component hierarchy, zones of interest, structure compression and enumerations of alternate configurations [3]. In effect, visual or text-based queries can be applied to the GPS structure, to extract representations for analysis on the cylinder surface. Design decisions are then made based upon multiple domain views emanating from the GPS. Each of these may be chosen to suit the domain expertise of the people viewing the information. A product centric internal representation will suit product designers, while process models are suitable for manufacturing management personnel.

Relationship information proceeds from the outer surface, and recedes into the centre, being deprecated/ escalated as required by the user. The central representation adapts in structure to eliminate clutter and pushes the relevant selected information to the outer cylinder. The path through the structure can be highlighted assisting traversal through the centre of the cylinder (see the red arrows in Fig. 2 to 5). Analysis using such a structure is kept to the surface of the outer cylinder, where a number of representations can be placed to support product process analysis. The internal representation provides a visual, summarized representation of the complex GPS hierarchies present in the PLM data, and thus allows the viewers to relate different components of the GPS to each other visually, while maintaining a sense of its whole structure.

IV. GPS HIERARCHY STRUCTURE

We now describe in detail how the internal GPS organizes itself with regards to user queries. We divide the GPS views up into a hierarchy, with roles determining views at the highest level. Deep hierarchies are reduced due to pragmatics as a form of structure compression that facilitates insight and speedier traversals. Thus the techniques use by

Eichhorn et al. [9] are integrated into the visualization, with an enforcement of role levels onto planes, overcoming issues with the complexity of the hierarchy and its interaction with the process models at lower levels.

In addition, each representation can be filtered either via a direct semantic querying, or by ordering in the natural hierarchy of the components. Each level "group" can be related to, in order from top down:

- Domains – as defined previously in Pools and Roles [9],[6];
- Views - related to the previous roles, derived from [10];
- Processes - and their respective perspectives [11], [12], 13];
- Product Architectures - levels of hierarchies in structure of the product [14];
- Components - base level to complex assemblies can be added to levels [14].

A. *Level Membership*

When displaying the GPS hierarchy, some of the major issues to address are the choice of the number of levels and their summarized contents. Due to the large data sets involved in a PLM GPS, this level membership becomes a complex issue of summarization [15] and selection [8] that will assist in reducing the complexity of the internal structures, and will help with interacting with the complex database present in the GPS. Two key factors in this presentation process are, the visual representations and the visual interactions provided by the adaptive level management.

The summarization process is intended to be adaptive in granularity. Objects and linkages can be adaptively summarized for each level to collapse complex multi-level hierarchies into simpler forms to populate each level [15]. This summarization process, however, will differ depending on the level being processed. Some levels in the GPS, such as the role hierarchies, are orders of magnitude less in complexity than say the deep and wide component hierarchies composing the product itself; this is especially due to people having responsibilities for more than one product or one component of a product. The number of levels visualized will be constrained to the main levels identified: Domains, Views, Processes, Product and Components. Each layer can be expanded and contracted as required, using an interactive animation.

Thus, the level membership is adaptive, but constrained to be within major organizational levels. The divisions are planar in nature, in order to promote a parallax effect when analyzing the structure of the GPS. For example, while a product hierarchy can be expanded to different levels across its tree, the hierarchy is constrained to the product level. These level hierarchies are then visualized within each level of the hierarchy using the previously mentioned Tree-Map technique. If such planar constraints are not implemented, then a feasible clear summary of the entire GPS would be problematic, due to imbalanced trees in the data model not being projected correctly onto each level plane, preventing parallax effects from providing insights into structures [8].

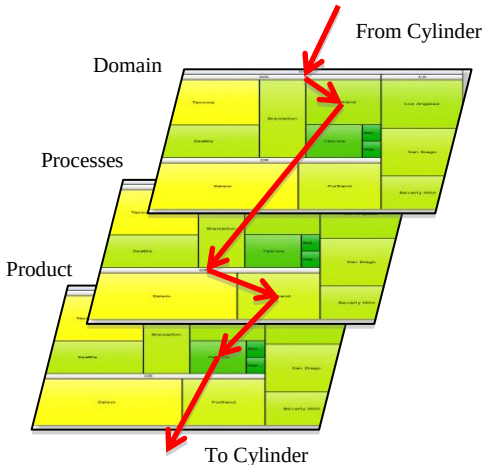


Figure 7: Hierarchy of Information Layers in the GPS.

In this simplified example, we show Domains, Processes, and Product as a selection of layers. The red arrows in Fig. 7 illustrate an example path through the hierarchy that indicates the context of the information presented on the outer cylinder, i.e., this is a detailed example of the coarsely illustrated red arrows shown in Fig. 2-7).

B. *Scalability Issues*

Tree-Maps are designed for large scale hierarchy representations, and so other methods for level selection and summarization can be exploited easily. The Product Hierarchy representation is simpler to process, due to its inherently hierarchical, non-planar, nature. Indeed, products and components lend themselves to such spatial 3D representations, as the manufactured products will have that structure on production, providing intuitive insight into the components under examination.

Product hierarchies can be summarized via the use of selective components in the product; that is, the choice of subcomponents as the key representation, such as a suspension unit, within an entire car. In addition, the summarization can be performed using selective exploded views on the complex subassemblies [14]; which can be intuitively performed via click or touch-based interactions, or by searches based on component or product key terms, e.g., strut, spring and/or cylinder.

Another issue to address with the PLM structure will be the maintenance of a sensible path through the internal data structure, to link together the outer cylindrical analysis visualizations, with the GPS. Thus, any of the highlighting mechanisms must allow for unbroken paths through the hierarchies, in order to facilitate user understanding of the context of the information as it is presented from the GPS. We show in Fig. 7, how the Tree-Map stack allows a summarized view of the relationships of each outer cylinder representation to the internal GPS structure.

Smooth interactions with the internal GPS will require progressive rendering algorithms applied to the large

 ISBN: 978-1-61208-181-6

hierarchy structures, in order to promote high levels of real-time interactivity. Animations of the modification of the environment will allow the extraction of relevant product representations, and will enable the visual tracking of changes by users in the structure of the GPS.

V. CONCLUSION

Tight production process collaboration between experts of different domains is required for successful operational decision processes. To improve this decision process, we have developed a conceptual framework, which enables a sophisticated visualization of multi-domain information in an engineering environment. Our framework should support the experts in their decision process by giving a holistic overview of the different domains, showing the impacts of changes in one domain to other domains. Although our framework is generic for multiple dimensions, we chose as examples the three domains: resource planning, manufacturing process planning and product construction.

Our framework is intended to be process oriented, due to a lack of domain independent process model representations in PLM. We connect the different domains to the process model domain. Therefore we aim to evaluate three modeling notations, Petri Nets, BPMN and YAWL as to their suitability to relate them to the resource and manufacturing domain.

We based our visualization approach on information retrieval mechanisms provided by PLM Systems. The items contained in a PLM can be assembled into a Global Product Structure. Using the GPS, views can be generated, according to different domain specialties. Based on these views, we developed our new 3D production process model representation. For the 3D presentation we proposed to use a cylinder structure. With this structure, it is possible for different domain experts to stand around the cylinder and view domain relevant analytic representations.

We discussed, and showed as mock-ups of two different representations for the inner structure (Product Centric, Tree-Map Stack). With these structures it is also possible to show the connection between the elements of the different views by highlighted assisted traversal through the center of the displayed data.

Future work will involve developing the system further, so it can also display the inner cylinder and show the connections between the elements of the different views in a fully interactive manner. Afterwards, we plan to perform an evaluation with domain experts about the usability of our visualization approach.

VI. REFERENCES

[1] H. Grabowski, R. Anderl, A. Polly, "Integriertes Produktmodell" 1.Auflage, Beuth Verlag, Berlin, 1993.

[2] Stark, John. Product Lifecycle Management: Paradigm for 21st Century Product Realisation, Springer-Verlag, London, 2004.

[3] A. Hambruegge, "Development of a method for evaluating the applicability of product views on a global product structure of an internationalautomotive manufacturer", M.Sc. thesis, KIT, Karlruhe, 2010.

[4] H. Stachowiak, „Allgemeine Modelltheorie“, Springer Verlag, Wien, New York, 1973.

[5] EN ISO 9001:2008, 2008.

[6] Effinger, J. Spielmann, Lifting business process diagrams to 2.5 dimensions, in: J. Park, M.C. Hao, P.C. Wong, C. Chen (Eds.), SPIE-IS&T Electronic Imaging: Visualization and Data Analysis, p. 75300O-75300O-9, 2010.

[7] U. Brandes, T. Dwyer, F. Schreiber. "Visual understanding of metabolic pathways across organisms using layout in two and a half dimensions", Journal of Integrative Bioinformatics, 2, 2004.

[8] B. Shneiderman, Tree visualization with tree-maps: 2-d space-filling approach, ACM Transactions On Graphics. 11 pp. 92-99, 1992.

[9] D. Eichhorn, A. Koschmider, Y. Li, A. Oberweis, P. Stürzel, R. Trunko, "3D Support for Business Process Simulation", Sheikh Iqbal Ahamed, Elisa Bertino, Carl K. Chang, Vladimir Getov, Lin Liu, Hua Ming, Rajesh Subramanyan, 33rd Annual IEEE International Computer Software and Applications Conference, IEEEComputer Society, Seattle,Washington, 2009.

[10] R. Bobrik, M. Reichert, T. Bauer, “Business process visualization-use cases, challenges, solutions”, Proceedings of the 5th international conference on Business process management Springer-Verlag Berlin, Heidelberg pp. 88—95, BPM'07, 2007.

[11] R. Brown, "Conceptual Modelling in 3D Virtual Worlds for Process Communication”, Asia-Pacific Conferences on Conceptual Modelling (APCCM 2010), Queensland University of Technology, Brisbane, 18-21 Jan. 2010.

[12] D. Eichhorn, A. Koschmider, "3D-Darstellung von Ressourcen-attributen bei der Geschäftsprozessmodellierung", Christian Gierds and Jan Sürmeli, Proceedings of the 2nd Central-European Workshop on Services and their Composition, CEUR Workshop Proceedings, CEUR.org, Berlin, Feb. 2010.

[13] M. Böhlen, L. Bukauskas P. Eriksen, S. Lilholt, L. Art,”3D Visual Data Mining—Goals and Experiences”, In Journal Computational Statistics & Data Analysis, of the International Association of Statistical Computing,43, pages 445–469, Elsevier Science, 2003.

[14] Z. Qiu, K. Kok, Y. Wong, J. Fuh, Role-based 3D visualisation for asynchronous PLM collaboration, Computers In Industry. 58 (2007) 747-755.

[15] A. Streit, B.L. Pham, R. Brown, Ross, "Visualization Support for Managing Large Business Process Specifications." In van der Aalst, Wil M. P., Benatallah, Boualem, Casati, Fabio, & Curbera, Francisco (Eds.) 3rd International Conference, Business Process Management, BPM2005, Nancy, France, 2005.

[16] T. N. Liukkonen, “VIPROSA – Game-like Tool for Visual Process Simulation and Analysis.” In Design and Use of Serious Games Intelligent Systems, Control and Automation: Science and Engineering, 2009, Volume 37, IV, 185-206, DOI: 10.1007/978-1-4020-9496-5_13.

[17] E. Kindler, C. Psáles, “3D-Visualization of Petri Net Models: Concept and Realization.” In Applications and Theory of Petri Nets 2004, Lecture Notes in Computer Science, 2004, Volume 3099/2004, 464-473, DOI: 10.1007/978-3-540-27793-4_27.

[18] J. J. Thomas and K. A. Cook (Ed.) (2005). “Illuminating the Path: The R&D Agenda for Visual Analytics” Chp. 3, National Visualization and Analytics, IEEE Computer Society, 2005.

 ISBN: 978-1-61208-181-6

Mastering Security Anomalies in Virtualized Computing Environments via Complex Event Processing

Lars Baumgärtner, Pablo Graubner, Matthias Leinweber, Roland Schwarzkopf, Matthias Schmidt, Bernhard Seeger, Bernd Freisleben
Department of Mathematics and Computer Science, University of Marburg
Hans-Meerwein-Str. 3, D-35032 Marburg, Germany
{lbaumgaertner,graubner,leinweberm,rschwarzkopf,schmidtm,seeger,freisleb}@informatik.uni-marburg.de

***Abstract*— To protect computer systems and their users against security attacks, all potential security related incidents should be detected by monitoring system behavior. In this paper, a novel approach to detect, analyze and handle security anomalies in virtualized computing systems is presented. Adequate sensors on different virtualization layers monitor relevant events, a Complex Event Processing engine is used to aggregate and correlate events on the same or different layers to find genuine attacks and eliminate false positives, and corresponding actions are performed if a security anomaly is detected. To enhance the quality of the results, machine learning techniques are used to analyze a historical database of recorded events offline to generate new or modify existing queries on the monitored event stream automatically. Furthermore, sensors can be activated and deactivated during runtime to gather interesting events, reduce the false alarm rate and ensure the system's responsiveness when a sudden increase of monitored event data occurs. In this way, a flexible, minimally-invasive approach for detecting, analyzing and reacting to a broad variety of security anomalies in a virtualized environment is provided.**

***Index Terms*— security; malware; virtual machine monitoring; complex event processing; intrusion detection.**

I. Introduction

Computers exposed to the Internet are at constant risk of being attacked. To protect them against security attacks, all security related incidents should be detected by monitoring system behavior. To detect security anomalies, Intrusion Detection Systems (IDS) or Intrusion Prevention Systems (IPS) are typically used; their combination is known as Security Information and Event Management (SIEM). However, most SIEM systems only monitor events on the infrastructural layer, need human assistance in case of error recovery, raise a high number of false alarms, and do not scale well with an increasing number of events.

In this paper, a new approach to detect, analyze and handle security anomalies is presented. The anomalies include both known and yet unknown security vulnerabilities with a particular focus on systems based on operating system virtualization, such as Infrastructure-as-a-Service Cloud computing systems. Hence, this paper proposes a novel SIEM system especially for virtualized computing resources.

The proposed architecture monitors security anomalies on different layers of a virtualized computing system. Monitoring is based on installing adequate sensors in the hypervisor (also called virtual machine monitor), in a virtual machine itself and in any kind of application runtime environment, such as an web-application container, to continuously report all relevant activities. A combination of out-of-VM monitoring using virtual machine introspection [6] and in-VM monitoring is used to keep the usual monitoring overhead low. To facilitate horizontal and vertical correlation and aggregation of monitored events, Complex Event Processing (CEP) is used. CEP enables robust cross-layer monitoring by leveraging Event Processing Agents (EPAs). EPAs are continuous queries on event streams that are able to analyze basic events and look for security anomalies. Based on the gathered results, the system can react autonomously and intelligently, i.e., it is able to repel attacks and circumvent security vulnerabilities. Occurring anomalies, even on different layers, will be detected at an early state and appropriate actions are launched. To enhance the quality of the results, machine learning is used to analyze archived, offline data. This allows us to generate new EPAs automatically through behavioral models derived from a historical database of recorded events. Furthermore, it is possible to activate and deactivate sensors during runtime in order to gather interesting, individual events, eliminate false positives and keep the system responsive when a sudden increase of monitoring data occurs. n this way, a flexible, minimally-invasive approach for detecting, analyzing and reacting to a broad variety of security anomalies in a virtualized environment is provided.

The core functionality is maintained in a special, trusted virtual machine (called ACCEPT-VM). It receives and processes all sensor data and triggers actions if necessary. In order to increase the processing speed and handle a larger amount of events, the benefits of multi-core architectures are leveraged, thus, EPAs can be scheduled between CPU cores. Furthermore, to make use of intra-EPA-parallelism, it should be possible to offload certain EPAs to General Purpose Graphic Processing Units (GPGPUs).

This paper is organized as follows. Section II discusses related work. Section III presents the approach for detecting, analyzing and handling security anomalies in virtualized computing systems. Section IV presents examples. Section V concludes the paper and outlines areas for future work.

 ISBN: 978-1-61208-181-6

II. Related Work

Teixera et al. [11] present Holmes, an implementation of a monitoring solution for integrating a CEP engine with machine learning. The CEP engine generates alerts using hand-crafted continuous queries to detect known abnormalities and deviations from expected behavior. Furthermore, it normalizes the asynchronous events for analysis with the machine learning algorithm, i.e., it joins different streams to be analyzed together and generates time series with equidistant intervals. A machine learning algorithm detects unknown anomalies in time series, without manual rule creation and anticipation of problem conditions and thresholds.

Holmes utilizes infrastructure level sensors and can thus only detect hard- and software issues as well as attacks such as Distributed Denial-of-Service (DDoS) attacks. The proposed architecture is not hierarchical, i.e., it consists of a single message bus, where all sensors publish their information to and the central CEP engine and machine learning modules subscribe to. This architecture does not scale well, neither for an increasing number of events nor for an increasing number of machines to monitor. Historical data is not used for anomaly detection, which limits the potential to detect anomalies as well as increases the risk of false positives.

Ficco [5] presents an approach to detect and respond to attacks by using event correlation. The approach is described using a DDoS attack as an example. Different information sources on several architectural levels, such as network, operating system and application, are deployed in strategic points of the system. In the example, these sources are the number of connections from a single IP, the length of the backlog queue of TCP and the number of application requests. Agents deployed together with the sensors analyze, filter, normalize and forward messages to the so called Decision Engine, consisting of a correlator, a diagnoser and a reaction module. Specialized modules, called Remediators, are used to remediate a specific attack or intrusion. An ontology is used to map all symptoms and possible effects of an attack. This ontology is used for the correlation of events and the decision about the right remediation strategy.

Although the proposed solution uses sensors on several architectural levels, it is targeted mainly at detecting different types of DoS attacks. The detection is based on the information about known attacks stored in the ontology. Detection of unknown anomalies is not possible with this solution. Since a central decision engine is used, scalability is also a problem of the architecture for growing network size or an increasing number of events. Finally, historical data is not taken into account in the detection process, missing another opportunity to eliminate false positives.

Cugola and Margara [3] present research about low latency CEP and general purpose GPUs. The work is based on the T-Rex CEP Engine and TESLA [4] as the language for defining rules. The authors assume that there are two major approaches for complex event processing: an automaton and a column based approach. Their main goal is to evaluate performance differences between them. Furthermore, they additionally compare them with a GPU variant. Automaton-based Incremental Processing (AIP) is the algorithm used to translate a CEP rule into a linear, deterministic finite state machine, which is fed with the incoming events while temporal results are stored. The counterpart approach is based on Column-based Delayed Processing (CDP), where events become more or less replicated for each rule and are stored in a column based structure. This algorithm is also used as a basis for their GPU implementation.

In all of their test setups, it is obvious that a CDP approach outperforms automaton based processing. They also implemented the CDP approach with CUDA on nVidia Graphics cards. Their first test of the GPU implementation results in a speedup of 25 compared with CPU CDP. Their evaluation leads to an average speedup of 40 with their hardware configuration. Additionally, their results are very varying depending on the particular configuration. For example, a large window size in which events are aggregated can lead to a speedup of 100. Their closing recommendation is to use GPU aided CEP only for large and complex rules, because there is a tradeoff between speedup and the overhead generated by using this technology.

Gorton [7] argues that the usage of a diversity of sensors on several architectural levels raises the chance to detect an attack, because the sensors may reinforce each other. However, this requires to manage and correlate the higher number of events and alerts. Different solutions have been developed in the area of intrusion correlation, targeted to the reduction of alerts a security officer must address. The potential to detect anomalies using these different information sources, however, is not the focus of these solutions.

III. Proposed Architecture

To detect attacks in a virtualized computing environment, it is useful to know as many anomalies in the behavior of the system as possible. Theoretically, every event that occurs in one of the different layers of a virtualized system can be an indicator for an anomaly: for example, established network connections, creation or termination of processes or even user or process activities beyond regular working hours.

The decision about what is a normal or unknown system behavior cannot be made by the sensors of a monitored environment. Instead, a CEP engine is responsible for processing all the informations sent by the sensors and is then able to decide what can be viewed as normal system behavior. Through dynamic deployment of further sensors, it is possible to eliminate false positives and verify findings.

Therefore, the architecture of the anomaly management system consists of a secure and trusted virtual machine (called ACCEPT-VM), where the main analysis components of the system such as the Sensor Management, the Matchmaker, the CEP engine and various databases are located. Furthermore, a set of passive sensors and actors reside on every layer (hypervisor, operating system and application layer) of each virtual machine. All of these sensors continuously deliver a stream of information to the ACCEPT-VM, and each actor is able to execute a specific set of actions on its corresponding layer in order to respond to any detected problem.

 ISBN: 978-1-61208-181-6

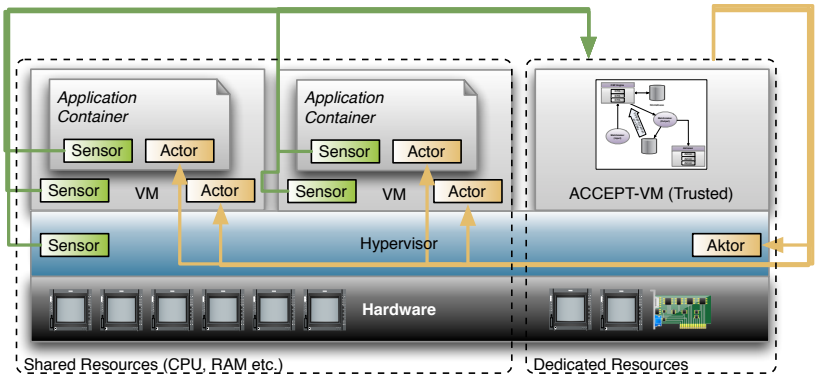


Fig. 1. Sensors and Actors in a monitored environment.

A. Monitored System

Sensors are deployed on several layers: The hypervisor, the operating system and application containers (see Figure 1). All sensors and actors are equipped with their required privileges related to the corresponding layer, and their implementation is aimed to be minimally invasive with respect to the normal functionality.

All information sent by sensors and received by actors contain important system information, which can be misused by an intruder. This is the reason why both the communication between sensors and the secure and trusted ACCEPT-VM as well as the communication between the ACCEPT-VM and the actors are secured in terms of authentication of the communication partners and the integrity/secrecy of the messages sent via the communication channels.

1) Sensors: There are several possibly interesting metrics to be gathered on each layer. Some of the expected events to be gathered can be found in the following list:

- Hypervisor level: Network traffic, system calls from within VMs, process lifecycle information.
- Operating system level: File access, network sockets, resource utilization.
- Application level: JVM information (heap utilization, thread count, library calls)

Sensors monitor traditional characteristics of resource usage (e.g., percentage of processor usage or memory consumption) as well as all data created by all processes in a system, such as system calls, network traffic, read/write memory access. This low level information greatly enhances the chances to detect more sophisticated attacks such as malicious polymorphic code, hidden processes or ongoing memory corruption exploits. On the hypervisor level, virtual machine introspection is used for acquiring monitoring data; on the operating system and application container level, the sensors are running as privileged user processes. Their security, as well as the security of the actors are ensured by hypervisor introspection. The hypervisor can monitor the running sensors and their process memory to guarantee that they have not been manipulated.

Sensors installed on the application level are used to monitor application behavior. This could be an application container such as Tomcat or JBoss or a bare Java Virtual Machine (JVM). Metrics gathered by these sensors are, e.g., changes of the memory heap, number of threads, number of Java classes, libraries or garbage collector statistics. Events occur if the code flow accesses constructors, methods and variables.

2) Actors: Actors are also installed across all layers, enabling direct countermeasures at the appropriate levels. Examples for such actions can be found in the following list:

- Hypervisor level: Start, stop or pause a virtual machine. Block or shutdown network interfaces.
- Operating system level: Start, stop, terminate processes or network connections. Delete users or files.
- Application level: Launch the garbage collector, solve deadlocks. Relaunch, terminate the application container or even remove components from the latter.

Further actions include a migration of a compromised virtual machine from the productive network to a separate honeypot network in order to detect possible malware. Since actions can be triggered on all layers of the virtualized system, they must be specified in a flexible, multi-purpose way. Therefore, an easy-to-use scripting language such as JavaScript is used. Actions can be executed concurrently on a target system, with the constraint of being executed in isolation, consistently and completely to avoid interferences or unknown system behavior. To increase the expressiveness and the usability of specified actions, actions are able to view all the data monitored by the sensor of the corresponding level.

B. ACCEPT-VM

The ACCEPT-VM consists of the CEP analysis engine, a Matchmaker, a Sensor Management, a Model Database and a Historical Database (see Figure 2). The latter is stored on a dedicated server in a data warehouse. The ACCEPT-VM is a trusted virtual machine in the sense that its attack surface is minimized: The number of services is reduced to an essential set, Mandatory Access Control is implemented and integrity checks are run on the file system. Existing security approaches, such as AppArmor [2] or SELinux [8], Tripwire [13] or AIDE [1] are used. Communication with this special virtual machine avoids direct TCP/IP traffic to the other virtual machines. A virtual device within the hypervisor is responsible for all message passing between sensors and actors. This has the advantage that the ACCEPT-VM cannot be directly attacked from the network or through a compromised host.

 ISBN: 978-1-61208-181-6

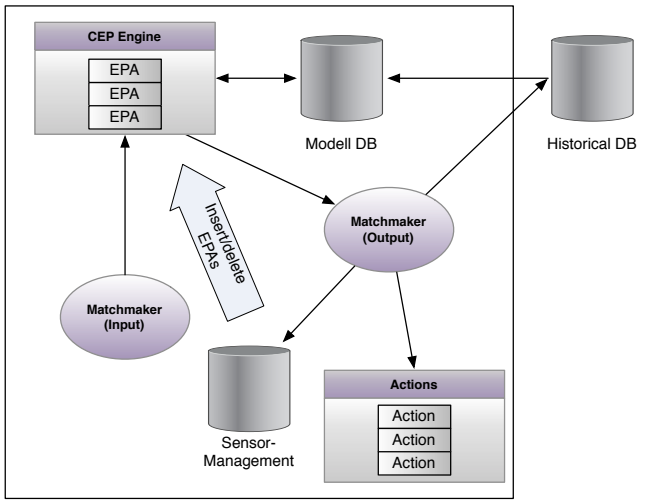


Fig. 2. Architecture of the ACCEPT-VM.

Furthermore, it provides a faster interface to pass information than classic socket communication.

The capability of performing analysis on a large amount of data on the input stream coming from the sensors is ensured by two main approaches: First, the EPAs running in the CEP engine are supported by pattern matching techniques. Second, the inherent control- and data-parallelism of EPAs are used to increase the number of events that can be processed. Furthermore, today's GPGPU technology is used to process a higher degree of events in parallel.

1) Complex Event Processing Engine: The analysis of all occurring events is performed using the CEP engine developed by the Software AG [10]. Event Processing Agents work as continuous queries on the different event streams coming from the sensors. These EPAs are verbalized in a SQL-like language and can combine information from different streams, thus different layers. Even mixing dynamic stream based queries with static data gathered, for example, from the Model Database (see III-B.2) is possible. Based on the results of the EPAs, actions are launched in order to react to a detected anomaly.

2) Model Database and Historical Database: The Model Database is built upon the information provided by the Historical Database. The Historical Database acts as a giant data store and foundation for model generation. It stores all events and actions generated by the sensors and the EPAs. Due to the high number of events generated by the sensors, it is necessary to develop new strategies to reduce the size of data. Nevertheless, it is necessary that the events and actions get recorded so that a root cause analysis is possible. The historical data can also be used to create simulations within a sandbox and replay certain scenarios. This enables both "what-if" analyses and the possibility to evaluate the effectiveness of EPAs.

With the data contained in the Historical Database, the Model Database can be created. This database contains the default behavior of a computer system expressed in statistical models. For example, a model regarding the average time of a TCP session to a web server can be created, and an EPA can be installed to detect abnormally long TCP sessions. Such a long open session might indicate a successful penetration of the web server with a shell running through the socket instead of short HTTP requests. The data from the Model Database can be used by the EPAs to detect essential anomalies differing from regular system behavior which again can trigger adequate actions.

3) Matchmaker: The Matchmaker component consists of two parts: The Input Matchmaker and the Output Matchmaker. The Input Matchmaker is responsible for a fully automatic interconnection between a sensor and an EPA. For example, if an EPA wants a specific input, then a request is sent to the Input Matchmaker. The Input Matchmaker is aware of the position of every sensor, including its meta-data, i.e., a flexible description of sensors and their possible actions. Analogously, the Output Matchmaker serves as a mediator that forwards EPA generated actions to actors within the monitored system. The effectiveness of these actions is measured by the Output Matchmaker. For example, if a false positive is detected, then the Sensor Management can take countermeasures by a reconfiguration of the corresponding EPA; if a static program analysis implies that a sensor is no longer useful, then it is possible to remove it from the system by means of the Sensor Management.

4) Sensor Management: The Sensor Management controls the (de-)activation and placement of sensors in a monitored system. It is able to (de-)activate sensors on-demand and during runtime, as well as it is able to scale the degree of data-granularity sent by a sensor. For example, the sampling rate of a sensor can be adapted to the needs of its corresponding EPA. Another advantage of this dynamic management system is that the number of events transmitted can be adjusted to an optimal level with respect to the system resources available to the ACCEPT-VM.

C. Performance Considerations

Due to the enormous number of sensors, a number of performance considerations have to be taken into account. Regarding the CEP engine, it is possible to optimize an EPA in order to perform multiple computations only once. While this problem is solved for simple queries, it still remains unsolved for complex ones. Hence, new methods for pattern matching must be developed. Furthermore, EPAs can be distributed to multiple, distributed and dedicated computing resources. If every EPA runs as a separate thread, it is possible to leverage the advantages of recent multi-core architecture to achieve a massive speedup (intra-EPA parallelism). It is also possible to execute the ACCEPT-VM on multiple, dedicated physical machines or an a whole compute cluster, respectively. It is likely that the system will generate huge amounts of data during runtime. Thus, the need for efficient ways to transfer events arises, both within a single virtualized node and between physical nodes in a virtualized cluster. The first problem can be solved by a lightweight secure communication channel based on para-virtualization. However, for inter-node communication, an approach is needed that reduces the amount of data, which could be achieved by technologies such as difference transmission.

 ISBN: 978-1-61208-181-6

GPGPU Aided CEP Engine: Due to the huge amount of data and probably very complex requests to detect sophisticated anomalies, it is necessary to take further technologies into account. A possible solution is the use of General Purpose Graphical Processing Units (GPGPU) with OpenCL or CUDA. Due to the architecture of this hardware, a high degree of parallelism can be achieved. But instead of a general use, only special kind of requests should be executed with this SIMD architecture due to hardware restrictions. It is imaginable that the CEP engines uses the GPGPU as a co-processor. Especially highly computation intensive tasks such as pattern matching can be outsourced with a great benefit, as indicated by the PFAC [9] library for exact string matching performed on GPUs.

If compression algorithms are required to dump the data to disk, a further application of the GPU might be possible. Due to the high speed stream processing capabilities of a GPU, it can be used to compress the event streams during runtime without generating additional load on the CPUs.

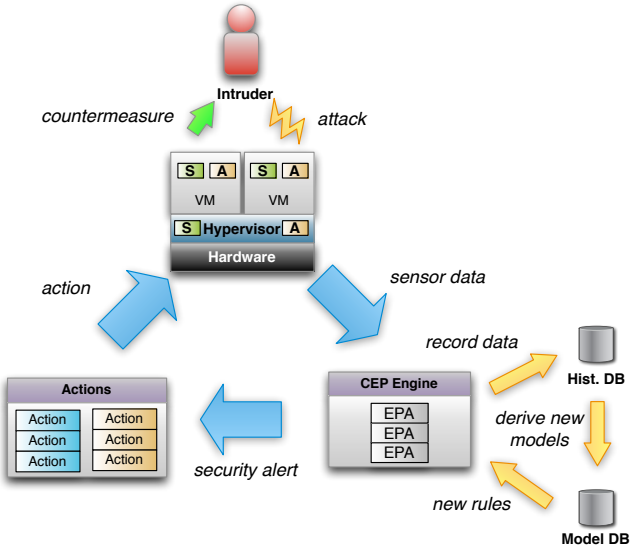


Fig. 3. ACCEPT Lifecycle

IV. Examples

The general lifecycle of the ACCEPT system is depicted in Figure 3. The ”blue” loop with the larger arrows represents the sense-detect-react-cycle and the ”yellow” loop with the smaller arrows shows the automatic rule generation process. To illustrate this new approach, two examples are presented in this section. First, a simple anomaly in double-entry accounting of the hypervisor and operating system layer port list is associated with a network based backdoor (see Subsection IV-A). Second, a more complex scenario shows an anomaly in the correlation of sensor data on the application container and operation system level (see Subsection IV-B), which is associated with a common attack scenario: An SQL injection attack [14].

A. TCP Backdoor

In this scenario, an attacker has successfully installed a backdoor in a monitored virtual machine. He/she hides his/her presence through a rootkit, a modification of the operating system and its userland interfaces. Even though the backdoor is listening on an arbitrary TCP port, the process belonging to it and the listening socket are not listed by the operating system userland tools.

Sensors: The scenario including the sensors and corresponding actions is shown in Figure 4. To detect the backdoor, at least two different sensors are involved: One sensor is running within the virtual machine and utilizes standard tools such as *netstat* to check for any listening sockets. Since the backdoor is well hidden, this sensor will not report the security breach.

The other sensor is inspecting the network state of the virtual machine from the hypervisor level. Since this sensor is running outside of the guest operating system, it is not affected by the backdoor’s hiding features. On this level, an event is generated for the detection of a newly opened port in the virtual machine.

Analysis: With the help of the Model Database and the Historical Database, queries can be generated to recognize normal or regular behavior. Therefore, an alarm should be triggered when a new open port is detected. Furthermore, by comparing both listening socket sensors, inside and outside of the virtual machine, it can be concluded that this really is a security related anomaly. A regular service installed in the virtual machine should not be hidden within the system. The conflicting sensors information is a clear sign of an attack.

Action: As a result of this attack, actions should be taken to eliminate the threat as much as possible. One such action could be to block all communication from and to the backdoor’s port on the hypervisor level. This prevents the attacker of extracting information or further using the infected machine. Another step that should be taken is to isolate and possibly terminate the processes involved in the infection. For forensics purposes, taking a snapshot of the virtual machine and generating a dump is another possibility.

B. SQL Injection

Another common attack scenario is the one of a web application vulnerable to SQL injection. Input not properly sanitized might lead to arbitrary queries executed on a backend SQL server. Even though SQL vulnerabilities could easily be fixed, they are still rated as the number one vulnerability in the Open Web Application Security Project Top 10 risk list [12].

Sensors: At least one sensor is needed to intercept all SQL statements sent from the web application to the corresponding SQL server. This sensor might be a JDBC proxy server or directly located in MySQL, for example. Furthermore, a sensor inspecting all running web applications and marking potentially harmful SQL statements. Furthermore, a sensor is needed to protocol which client request caused which SQL queries.

Analysis: All queries coming from potentially harmful marked statements can then be further analyzed. Signs of potential SQL injection attempts can be identified by various syntactic specialties such as extensive use of wildcard *SELECT*s or trimmed of code through comments. In conjunction

 ISBN: 978-1-61208-181-6

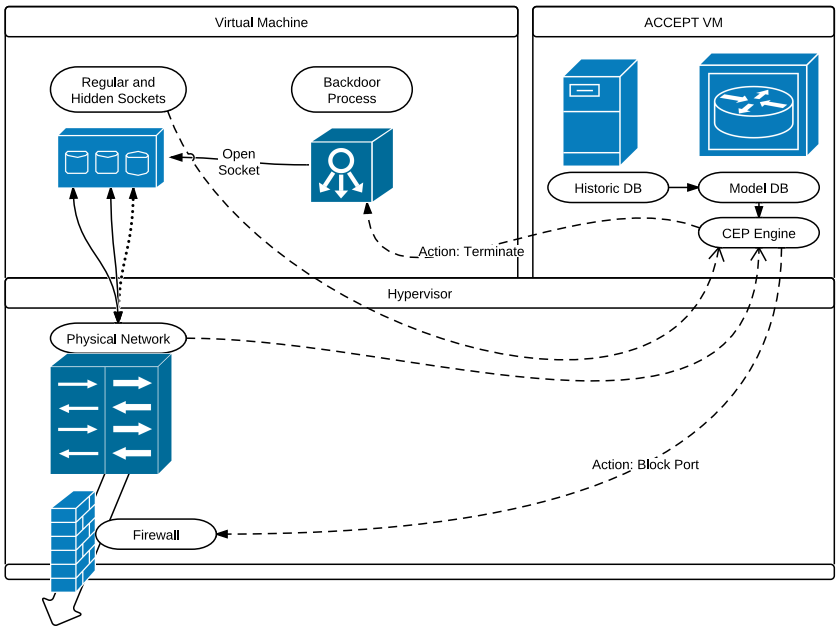


Fig. 4. Example: TCP backdoor detection.

with the information of the source of this request, an alarm can be raised. One by-product of the passive analysis is that potentially harmful code can be found before it is exploited and fixed in advance.

Action: The obvious action to take is blocking the offending IP address through a firewall. Through the use of the code inspection sensor, the faulty code sections might also be directly visible and automatic solutions to fix the problem can be used.

V. Conclusion

In this paper, a multi-level approach incorporating several state-of-the-art techniques such as virtualization, virtual machine introspection and complex event processing for detecting, analyzing and handling security anomalies has been presented. The proposed approach tries to keep management and maintenance within the virtual machines to a bare minimum by emphasizing the use of sensors on the hypervisor level. Furthermore, it does not only detect security anomalies, but is also able to react accordingly and defend or secure the system automatically. The flexible nature of the framework and the CEP backend make it especially easy to add new sensors to increase security and react to new threats or adapt to new technologies/devices. Using EPAs allows robust monitoring on all layers. With the Historical Database and techniques from machine learning, new EPAs can be automatically generated. While being able to detect not only new anomalies, the system can also verify false positives through correlation of different sensor layers. The components of the proposed framework form a reliable and adaptable security monitoring solution.

The approach is currently under development, and no significant implementation work has been started yet. Future work is devoted to detail the design of the components, implement the system and perform adequate experimental evaluations.

References

[1] AIDE Project. Advanced Intrusion Detection Environment (AIDE). http://aide.sourceforge.net/, 2011. retrieved: November, 2011.

[2] M. Bauer. Paranoid Penguin: An Introduction to Novell AppArmor. *Linux Journal*, 2006:13, August 2006.

[3] G. Cugola. Low Latency Complex Event Processing on Parallel Hardware. Technical report, Politecnico di Milano, 2011.

[4] G. Cugola and A. Margara. TESLA: A Formally Defined Event Specification Language. In *Proceedings of the Fourth ACM International Conference on Distributed Event-Based Systems*, DEBS '10, pages 50–61, New York, NY, USA, 2010. ACM.

[5] M. Ficco. Achieving Security by Intrusion-Tolerance Based on Event Correlation. *Network Protocols and Algorithms*, 2(3):70–84, 2010.

[6] T. Garfinkel and M. Rosenblum. A Virtual Machine Introspection Based Architecture for Intrusion Detection. *Proceedings of the 2003 Network and Distributed System Security Symposium*, pages 191—206, Jan 2003.

[7] D. Gorton. Extending Intrusion Detection with Alert Correlation and Intrusion Tolerance. *Licentiate Thesis, Chalmers University of Technology*, 2003.

[8] National Security Agency. Security-Enhanced Linux. http://www.nsa.gov/research/selinux/, 2009. retrieved: November, 2011.

[9] PFAC open library. PFAC Open Library for Exact String Matching Performed on NVIDIA GPUs. http://code.google.com/p/pfac/, 2011. retrieved: November, 2011.

[10] Software AG. webMethods Business Events. www.softwareag.com/corporate/products/wm/events/capabilities/default.asp, 2011. retrieved: November, 2011.

[11] P. H. S. Teixeira, R. G. Clemente, R. A. Kaiser, and D. A. Vieira Jr. HOLMES: An Event-driven Solution to Monitor Data Centers through Continuous Queries and Machine Learning. In *Proceedings of the Fourth ACM International Conference on Distributed Event-Based Systems*, pages 216–221. ACM, 2010.

[12] The Open Web Application Security Project. OWASP Top 10 Risks. https://www.owasp.org/index.php/Top_10_2010-A1, 2010. retrieved: November, 2011.

[13] Tripwire Inc. Tripwire. http://www.tripwire.com/, 2011. retrieved: November, 2011.

[14] W.G. Halfond J. and Viegas and A. Orso. A Classication of SQL-Injection Attacks and Countermeasures. In *In Proceedings of the International Symposium on Secure Software Engineering*, 2006.

Knowledge Discovery Using a Service Oriented Web Application

Janez Kranjc, Vid Podpečan, Nada Lavrač
Department of Knowledge Technologies
Jožef Stefan Institute
Ljubljana, Slovenia
{*janez.kranjc,vid.podpecan,nada.lavrac*}*@ijs.si*

***Abstract*—The paper proposes a novel platform for knowledge discovery, which is based on modern web technologies, and is implemented as a web application. It is based on the principles of service-oriented knowledge discovery, and features interactive scientific workflows. In contrast to the few existing comparable platforms, our environment is suitable for any knowledge discovery task, offers advanced workflow construction including meta-workflows, can use any existing web service as a workflow processing component, and runs in all major web browsers and platforms, including mobile devices. The presented environment has been demonstrated on two use cases: a simple motivating use case built using Weka web services, and an advanced use case featuring a complex text mining scenario.**

***Keywords*-data mining; knowledge discovery; web application; web services**

I. Introduction

The development of modern knowledge discovery and data mining environments and tools has reached maturity. While traditional data analysis software supported a single or few algorithms, designed to mine highly specialized data, today's modern knowledge discovery systems provide a large collection of generic algorithm implementations, usually coupled with an easy-to-use graphical user interface. The importance of visual programming using scientific workflows is now also widely recognized, and all advanced knowledge discovery software offer some form of workflow construction and execution, as this is of crucial importance for conducting complex scientific experiments, which need to be repeatable, and easy to verify at an abstract level.

However, these so-called second generation systems have failed to benefit from the concepts of service-oriented architecture, and complex and geographically dispersed information and knowledge sources as well as algorithms and functions, publicly available on the web. Finally, today's knowledge discovery systems have also failed to bridge different operating systems and platforms, and are not able to fully utilize available server resources in order to relieve the client from heavy-duty processing and data transfer. As the general trend is shifting towards mobile devices and mobile computing, this effectively prevents the employment of such tools in modern mobile information environments.

The novel knowledge discovery platform presented in this paper was designed to overcome all the recognized deficiencies while retaining all important features of existing solutions. As such, our platform benefits from service-oriented technologies [1], the visual programming paradigm, as well as platform *and* software independent technologies.

Firstly, service-oriented architecture featuring web services as primal processing components enables parallelization, remote execution, and high availability by default. It provides access to large public (and proprietary) databases, enables easy integration of third party components (as services) and loose coupling, and supports not only distributed processing but also distributed development.

Secondly, the visual programming paradigm simplifies the construction of complex knowledge discovery scenarios by providing basic building blocks, which can be connected and executed, enables repeatability of experiments by saving constructed workflows and parameters, provides an intuitive structuring of complex parts of workflows by introducing the notion of meta-workflow (workflow of workflows), and makes the platform suitable also for non-experts due to the representation of complex procedures as sequences of simple processing steps.

Finally, as the platform and software independence can be achieved by using web technologies only, the platform relies on standards such as HTML, CSS, Ajax and JavaScript, and widely supported and accepted software solutions such as Apache and PHP.

To summarize, the presented platform offers a complete service-oriented workflow environment, suitable for any knowledge discovery task. The platform is truly independent as it is implemented in the form of a web application, which is accessible from any modern web browser.

The rest of the paper is structured as follows. Section II briefly presents the related work. In Section III, the design of the platform and its components are discussed in detail. The description of the initial widgets repository is presented in Section IV. In Section V, two data mining use cases are presented. Finally, Section VI summarizes the work and concludes the paper by suggesting directions of further work.

II. Related work

This section discusses the work related to the main concepts of the presented platform: workflow editing and execution environments, service-oriented approaches to knowl-

 ISBN: 978-1-61208-181-6

edge discovery and browser-based applications for knowledge discovery.

Many software solutions from different domains enable construction and execution of scientific workflows. Well known examples include Weka [2], Orange [3], KNIME [4] and RapidMiner [5] data mining environments. The most important common feature is the implementation of a *workflow canvas* where complex workflows can be constructed using simple drag, drop and connect operations on the available components. The range of available components typically includes data loading and pre-processing, data and pattern mining algorithms and interactive and non-interactive visualizations.

Even though such data mining software solutions are reasonably user-friendly and offer a wide range of components, there are many deficiencies, which limit their use. Firstly, all available workflow components provided by any of these platforms are specific and can be used in this particular platform only. Secondly, the described platforms are implemented as standalone applications and have specific hardware and software dependencies. Thirdly, in order to extend the range of available workflow components in any of these platforms, knowledge of a specific programming language is required. This also means that they are not capable of using existing software components, implemented as web services, and available freely on the internet.

In order to benefit from service-oriented architecture concepts, another group of software tools have emerged, which are able to make use of web services, and can access large public databases (some also support means of grid deployment and P2P computing). Environments such as Weka4WS [6], Orange4WS [7], Web Extension for RapidMiner, Triana [8], Taverna [9] and Kepler [10] allow for integration of web services as workflow components. However, with the exception of Orange4WS and Web Extension for RapidMiner, these environments are mostly specialized in domains like systems biology, chemistry, medical imaging, ecology and geology. Lastly, none of these applications is browser based thus still requiring specific hardware and software.

The last group of software tools capable of workflow construction, most similar to the presented environment, encompasses browser based applications such as the Oryx Editor [11] for modelling workflows and business processes, and the Galaxy [12] genome analysis tool. The Oryx editor, however, although designed similarly as the proposed environment, does not support the deployment of workflows as it is only a modelling tool. Also, the Galaxy web application is limited exclusively to the workflow components, provided by the project itself, and does not provide means for employing arbitrary web services and other information and computing resources found on the web.

III. Platform Design

The presented platform consists of three layers as shown in Figure 1. The upper-most layer presents the parts of the platform which run on the client side. The middle layer is located on the server where the platform is hosted. The bottom layer consists of the remote resources which provide web services.

This section describes these layers in detail. The user interface is presented first. Secondly, the workflow engine is described. Finally, workflow components and the functionalities to import web services are explained.

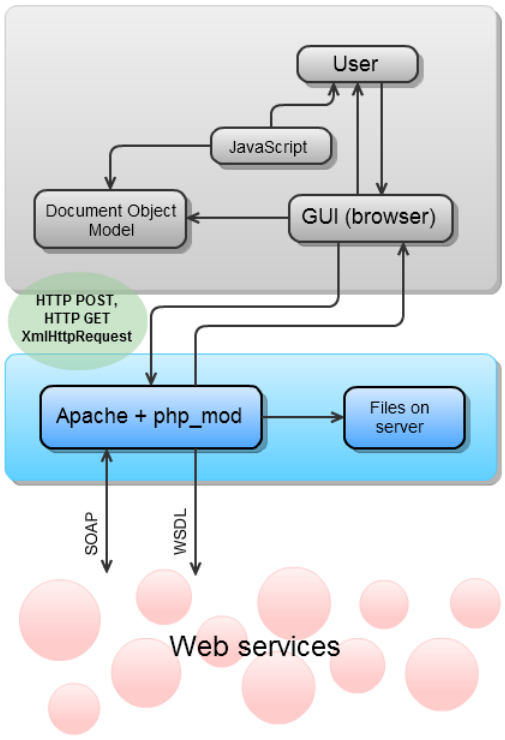


Figure 1. The three layered design of the platform. The upper-most layer represents the part of the platform executed on the client side. The middle layer is located on the server where the platform is hosted. The lower layer represents remote resources which provide web services.

A. The user interface

The graphical user interface was implemented in HTML. It consists of three main parts: the toolbar, the widget repository, and the canvas. A sample screenshot of the user interface is shown in Figure 2.

Two JavaScript libraries were used to implement the toolbar. The buttons were implemented using the jQuery UI library while their event listeners and handlers were implemented using the jQuery library [13]. The primary function of the toolbar is to start, execute, save, and load workflows, and to separate parts of the workflow.

 ISBN: 978-1-61208-181-6

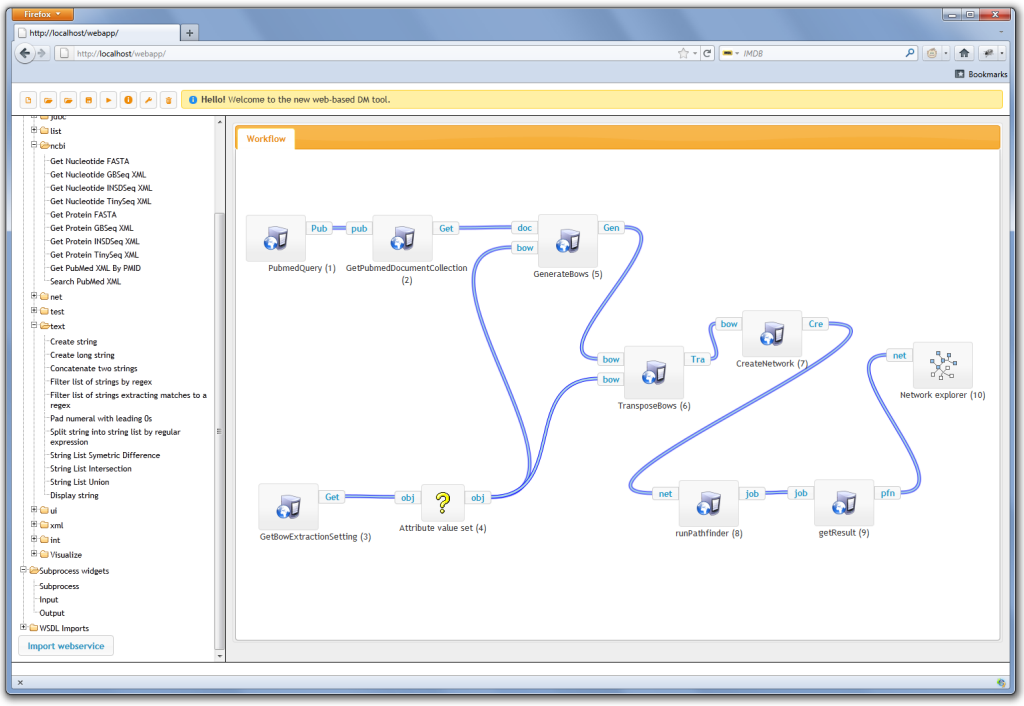


Figure 2. A screenshot of the environment in the Mozilla Firefox browser.

The second part of the graphical user interface is the widget repository, which provides a clickable list of available widgets. By clicking on an available widget, its instance appears on the canvas hosting the currently active workflow.

As the workflow canvas is the part which requires intensive user interaction, it was implemented in JavaScript. Each widget on the canvas is represented by a short HTML description, and a special function in JavaScript, which appends this HTML code to the document object model, is invoked whenever a widget from the repository list is clicked.

Widgets can be connected by clicking on an input or output. When both an input and an output are selected, an event is triggered, which checks for cycles in the workflow graph using the depth first search algorithm. If no cycles are detected, a connection is drawn and the corresponding widgets become connected.

The connections are graphically represented as Bézier curves, and implemented by dynamically adding HTML5 canvas elements to the document object model. To enable cross-browser functionality, the ExplorerCanvas library was used to simulate the canvas element in the Internet Explorer browser family. Finally, the selection and removal of connections are implemented using mouse and keyboard event handlers, respectively.

B. The workflow engine

The presented platform is able to execute workflows as well as separate widgets. A PHP script on the server corresponds to each widget. It is invoked from the graphical user interface using Ajax techniques.

The inputs of the widget are passed to the PHP script using an asynchronous HTTP POST request. When the results are available (or when an error occurs), a call-back function is called, which stores the results of the execution of the widget into the output variables in the underlying document object model. The PHP script may either return the data in a serialized form or issue a special command, which instructs the user interface to open a pop-up window for displaying the results (data visualization widgets utilize this functionality). The execution of multiple independent widgets simultaneously is assured by the asynchronous nature of POST requests, which essentially represent an equivalent to the multithreading programming paradigm.

The execution of the entire workflow is realized by a special JavaScript function, which iteratively searches for widgets whose predecessors have finished their execution, and executes them.

C. Web services as workflow components

Workflow components of the presented platform may be implemented as remote web services provided by a third party, or as PHP scripts located on the server hosting the platform.

 ISBN: 978-1-61208-181-6

Since web services are completely defined by their WSDL descriptions, the functionality to import web services was implemented in PHP by parsing the corresponding WSDL document. For every operation provided by a web service the PHP script returns an HTML description of the corresponding widget. In the user interface, this procedure is accessible through a button whose event handler queries the user for the location of a WSDL file, which is then imported and parsed, and a list of available widgets is presented to the user.

The user is allowed to decide about the role of each input of each operation. The input may be designated as a user interface input of the widget or as a widget input parameter. User interface inputs can be set by entering the values manually in the widget's graphical user interface while widget input parameters receive data from other widgets.

IV. The widget repository

In order to enable construction of scientific workflows implementing arbitrary knowledge discovery scenarios, an initial set of widgets is available to the user. The widgets belong to four distinct groups.

The first group of widgets enable data creation, manipulation and simple visualizations including creation of strings and integers, and joining strings and integers into arrays. Arithmetic integer operations are also implemented. These widgets are implemented as PHP scripts located on the server where the platform is running.

The second group of widgets consists of three widgets, which allow creating nested workflows (note that the depth is unlimited). A sub-workflow can be created by adding an instance of the sub-workflow widget on the canvas. On activation, the canvas view is switched to the sub-workflow (switching can also be achieved by clicking on the corresponding canvas tab). Additionally, two special widgets are used for assigning inputs and outputs of the current sub-workflow to carry the data from the parent workflow to the sub-workflow. By adding an input widget to the current sub-workflow, the corresponding sub-workflow widget gains an input, which is connected to the output of the input widget. The output widget operates similarly. This group of widgets is implemented in JavaScript and executed on the client side.

The third group of widgets consists of 35 implementations of the local services available in Taverna [9]. Because Taverna is written in Java and its local services in the Beanshell scripting language, they were implemented in PHP and integrated into the platform. These services include widgets that allow reading and writing files, re-encoding strings, executing SQL queries, querying public databases such as PubMed, accessing documents using HTTP, extracting and viewing images from websites, performing operations on strings and lists of strings, and manipulating XML files. Due to security issues in browsers, some services could not have been implemented in PHP, such as services that list files and folders of the user's computer and execute applications. Using the implemented collection of Taverna services our platform supports the majority of workflows created in Taverna.

Finally, the platform offers several data mining algorithm implementations from the Weka data mining environment, which have been made available as SOAP web services. The Orange4WS [7] data mining platform and its tools were used to implement these services, which enable easy and platform independent access to the latest Weka software. The actual SOAP web server hosting Weka services makes use of the JPype wrapper library [14], which allows calling arbitrary Java code from the Python interpreter. The services communicate by exchanging serialized Weka objects such as learners, classifiers, and datasets. While this approach does not allow client-side modification of these objects unless a Java client running Weka is used, it is currently the only feasible way to use Weka in a service-oriented environment as only a few Weka classes implement XML-based serialization according to PMML standard [15].

V. Use cases

This section demonstrates some of the abilities of the presented platform. Two use cases are presented and discussed. The first one is a simple, motivating use case where Weka web services are used to show the basic functionalities of the platform: composing a workflow of remote web services and local processing components, executing the workflow, and displaying the results. The second use case is an advanced example of a text mining workflow where a word graph obtained by querying a public database is pruned using a specialized graph algorithm (available as a web service) and visualized locally in a powerful interactive graph visualization component, provided by the platform.

A. A use case featuring Weka web services

The purpose of this use case is to demonstrate the use of Weka algorithms, available as web services (see Section IV). The J48 decision tree induction algorithm (J48) is cross-validated on the 1984 United States Congressional Voting Records dataset, which consists of 435 instances with 17 attributes and a binary class attribute, and is available in the Attribute-Relation File Format (ARFF).

First, the *Read Text File* widget is used to load the dataset into the platform. The widget provides a file chooser dialog where an arbitrary text file can be selected and uploaded to the server.

Since Weka web services communicate by exchanging serialized Weka objects, the dataset has to be transformed into a serialized Weka Instances object. A service from the repository of Weka services provides this functionality.

Then, to perform the cross-validation of a learning algorithm the Weka service for cross validation is used. It accepts serialized data, a serialized Weka learner (i.e., an

 ISBN: 978-1-61208-181-6

implementation of the learning algorithm), and the number of folds. The Weka service providing the J48 learner is connected to the cross-validation service, which can finally be executed. The cross-validation service provides several outputs, which return different reports generated by Weka. Here, Weka's cross validation summary is displayed using the *Display String* widget.

The complete workflow for performing cross validation using Weka services and displaying the results is shown in Figure 3. The workflow can be executed by clicking the execute workflow button. It invokes the workflow engine, which takes care of the validation, execution, and error reporting.

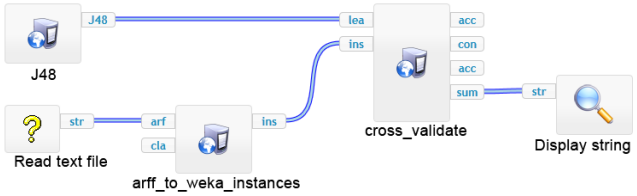


Figure 3. A workflow implementing cross validation of Weka algorithms.

B. An advanced text mining use case

This use case is built upon a collection of advanced services for text mining, graph analysis, and graph visualization. Its goal is to support the analysis of textual data by providing means for representing texts as graphs, graph pruning, and interactive graph visualization.

The information source for this use case is the well known PubMed database [16], a free database accessing the MEDLINE database of citations, abstracts and some full text articles on life sciences and biomedical topics. It was used for obtaining documents, relevant to the input query.

The resulting document corpus was then processed using text mining tools from the LATINO project [17], which is a software library implementing a range of data mining and machine learning algorithms with the emphasis on text mining and link analysis (components of the LATINO library components were provided as web services).

Using LATINO web services, the document corpus was transformed into a term network as follows. Firstly, it was tokenized and lemmatized, and transformed into the bag-of-words (BoW) vectors. Then, the network was produced using the following principle. Each link between two words represents a co-occurrence meaning that both words appear together in at least one document, whereas the link's weight represents the normalized number of co-occurrences through all documents. This weight is a similarity measure in the sense that two words (or concepts) linked with a higher weight are more similar (i.e. are more "connected" because they appear together more often) than two words with a lower weight.

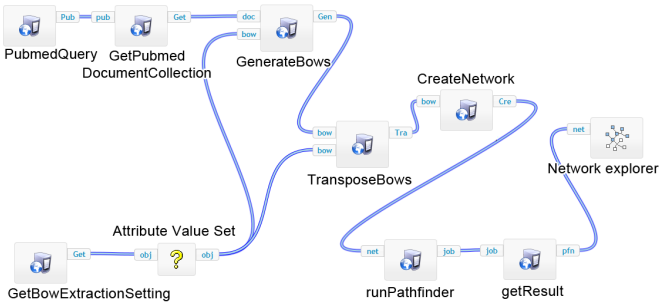


Figure 4. A workflow implementing the analysis of textual data using a public database and a collection of text mining and graph mining components.

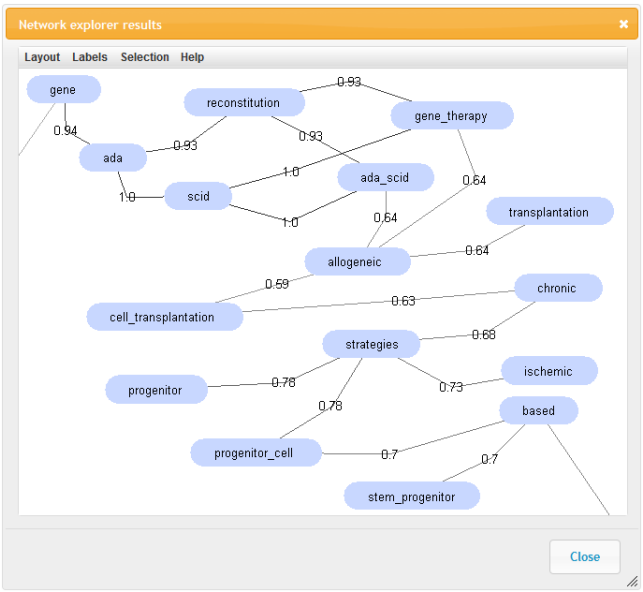


Figure 5. A part of the term graph obtained by querying PubMed with the query string *stem cell*. The graph is visualized using an interactive Java applet available as a widget in the presented platform.

Finally, the resulting weighted graph was pruned using the Pathfinder algorithm [18], a specialized algorithm for graph simplification (we omit the details of the algorithm as this is out of the scope of this paper). In order to transform the weights into dissimilarities required by Pathfinder the formula $w = 1/w'$ was applied where w is the original weight.

As a result, the pruned graph retained only the strongly linked concepts, which means that many less significant edges were removed thus improving its understandability and presentability. In the constructed workflow, as shown in Figure 4, an interactive graph visualization component was used, which enables user friendly exploration of large graphs. A zoomed part of the graph, obtained by querying PubMed with the query *"stem cell"*, is shown in Figure 5.

 ISBN: 978-1-61208-181-6

VI. Conclusion and futher work

The paper proposes a browser-based environment for service oriented knowledge discovery, which relies on modern web standards and widely supported and accepted software solutions. Coupled with the extreme versatility and power of web services, the proposed environment presents a new generation tool, ready to be used in any scenario or form of knowledge discovery, including mining of web and data streams thus surpassing all currently available knowledge discovery software tools. Moreover, the proposed environment is able to run in all modern web browsers, including those available on mobile devices, which presents great opportunities for its deployment and widespread use.

In summary, we have developed an open, general, and independent web application environment for knowledge discovery, which employs service-oriented technologies, and is ready to be used in any data and information analysis scenario.

In future, we plan to implement the process flow control widgets such as conditional branching and looping. We will also explore adding means of mining data streams as well as semi-automatic workflow construction based on planning algorithms, modern knowledge discovery ontologies, and systems for semantic annotation of web services. Finally, we plan to provide a public installation of the environment, a workflow repository, a community web site, and release the sources under an open-source license.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) within the context of the project FIRST, Large scale information extraction and integration infrastructure for supporting financial decision making, under grant agreement no. 257928.

References

[1] T. Erl, *Service-Oriented Architecture: Concepts, Technology, and Design*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2005.

[2] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Amsterdam: Morgan Kaufmann, 2011.

[3] J. Demšar, B. Zupan, G. Leban, and T. Curk, "Orange: From experimental machine learning to interactive data mining," in *PKDD*, ser. Lecture Notes in Computer Science, J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, Eds., vol. 3202. Springer, 2004, pp. 537–539.

[4] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, "KNIME: The Konstanz Information Miner," in *GfKl*, ser. Studies in Classification, Data Analysis, and Knowledge Organization, C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker, Eds. Springer, 2007, pp. 319–326.

[5] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler, "YALE: rapid prototyping for complex data mining tasks," in *KDD*, T. Eliassi-Rad, L. H. Ungar, M. Craven, and D. Gunopulos, Eds. ACM, 2006, pp. 935–940.

[6] D. Talia, P. Trunfio, and O. Verta, "Weka4WS: A WSRF-enabled Weka toolkit for distributed data mining on grids," in *PKDD*, ser. Lecture Notes in Computer Science, A. Jorge, L. Torgo, P. Brazdil, R. Camacho, and J. Gama, Eds., vol. 3721. Springer, 2005, pp. 309–320.

[7] V. Podpečan, M. Zemenova, and N. Lavrač, "Orange4ws environment for service-oriented data mining," *The Computer Journal*, 2011.

[8] I. Taylor, M. Shields, I. Wang, and A. Harrison, "The Triana workflow environment: Architecture and applications," *Workflows for e-Science*, vol. 1, pp. 320–339, 2007.

[9] D. Hull, K. Wolstencroft, R. Stevens, C. A. Goble, M. R. Pocock, P. Li, and T. Oinn, "Taverna: a tool for building and running workflows of services," *Nucleic Acids Research*, vol. 34, no. Web-Server-Issue, pp. 729–732, 2006.

[10] I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludascher, and S. Mock, "Kepler: an extensible system for design and execution of scientific workflows," in *Scientific and Statistical Database Management, 2004. Proceedings. 16th International Conference on*, Jun. 2004, pp. 423–424.

[11] G. Decker, H. Overdick, and M. Weske, "Oryx – an open modeling platform for the bpm community," in *Business Process Management*, ser. Lecture Notes in Computer Science, M. Dumas, M. Reichert, and M.-C. Shan, Eds. Springer Berlin / Heidelberg, 2008, vol. 5240, pp. 382–385.

[12] D. Blankenberg, G. V. Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, and J. Taylor, "Galaxy: A Web-Based Genome Analysis Tool for Experimentalists," *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, vol. Chapter 19, Jan. 2001.

[13] "jQuery: The write less, do more, javascript library," Last accessed January 2012. [Online]. Available: http://jquery.com

[14] "JPype - java to python integration," Last accessed January 2012. [Online]. Available: http://jpype.sourceforge.net/

[15] "PMML 4.0 - general structure of a pmml document," Last accessed January 2012. [Online]. Available: http://www.dmg.org/v4-0-1/GeneralStructure.html

[16] "Pubmed," Last accessed January 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/

[17] M. Grčar, "Latino - link analysis and text mining toolbox," Last accessed January 2012. [Online]. Available: http://sourceforge.net/projects/latino/

[18] A. Vavpetič, A. Batagelj, and V. Podpečan, "An implementation of the pathfinder algorithm for sparse networks and its application on text networks," in *Proceedings of the 11th International Multiconference Information Society*, 2009.

 ISBN: 978-1-61208-181-6

Self-Learning Monitoring and Control of Manufacturing Processes Based on Rule Induction and Event Processing

Daniel Metz, Sachin Karadgi, Ulf Müller, Manfred Grauer
Information Systems Institute,
University of Siegen,
Siegen, Germany
{Daniel.Metz, Sachin.Karadgi, Ulf.Mueller, Manfred.Grauer}@uni-siegen.de

Abstract **- Manufacturing enterprises are trying to cope with turbulent market situations by enhancing their existing monitoring and control of manufacturing processes. Enterprise integration within and across the enterprise can assist to realize the aforementioned goal. Further, event processing (EP) techniques can be employed to monitor and control manufacturing processes in real-time. Rules derived from stored process data using the knowledge discovery in databases process can be managed in an EP engine as event patterns. Nonetheless, rule identification is usually an offline activity whereas the control of manufacturing processes is a real-time activity. Consequently, the rule identification process should be transformed from an offline activity to an online or (near) real-time activity. In the contribution, a methodology is presented to overcome the previously mentioned drawback. Machine learning (i.e., rule induction) methods are used to automatically adapt the existing set of event patterns. The implementation of the presented methodology has been started in a casting enterprise.**

Keywords - complex event processing, rule induction, rule classification, knowledge management, real-time control, machine learning.

I. INTRODUCTION AND PROBLEM DESCRIPTION

Manufacturing enterprises are compelled to manufacture products with high quality and shortened lead times due to rapidly changing customer requirements, decreased life-cycle of products, and drastic variation in environmental conditions. Further, these challenges are intensified for enterprises having a high product mix with low volume production. Usually, these enterprises operate monitoring and control systems for their manufacturing processes. Nevertheless, enterprises need to enhance their existing systems to monitor and control manufacturing processes to retain their competitive advantages. Especially, enterprises strive for a higher degree of flexibility and adaptability.

The main challenge on the route to achieve the aforementioned vision is the necessity for an integrated enterprise [1]. Attempts are being made to integrate enterprise levels along horizontal and vertical direction based on ISO 15704, enterprise reference architecture [2]. Horizontal integration deals with the integration of enterprise applications (e.g., enterprise resource planning (ERP) system) or resources at a particular enterprise level. On contrary, enterprise levels are associated with different time horizons, which require vertical integration of information and knowledge. Overall, enterprise integration (EI) within and across different enterprise levels can provide a holistic view of an enterprise. Therefore, EI can be considered as a prerequisite for enhancing the monitoring and control of manufacturing processes. Further, EI can be exploited to accomplish management strategies like the real-time enterprise (RTE), support enterprise performance measurement, and enhance knowledge management (KM), among others.

During execution of manufacturing processes, enormous amount of process data (e.g., sensor readings, product quality feedbacks) is generated in (near) real-time (i.e., seconds or milliseconds) by resources located on the shop floor. In addition, operators provide necessary data related to resources, orders and products (e.g., selecting pre-defined reasons for a resource breakdown, order details during start of an order execution). This process data is utilized in different ways. First, the data is displayed in (near) real-time to enterprise members for monitoring of manufacturing processes. Second, process data is stored in relational databases for offline analysis (e.g., deriving new knowledge). Finally, the data is employed for real-time monitoring and control of manufacturing processes based on event processing (EP).

EP has become an appropriate technology for event-driven applications [3]. The knowledge (i.e., rules), derived from the offline analysis of stored / historical data using analytical techniques (e.g., data mining), can be modeled as event patterns in an EP engine. The rules can also be obtained from structured interviews with domain experts. Additionally, reactive rules can be defined, which describe (re-) actions to situations detected by analyzing the incoming process data streams. However, rule identification is an offline activity whereas controlling of manufacturing processes is a real-time activity.

In addition, today's shop floor is characterized by high automation and few employees, an employee managing multiple resources. Accordingly, the control system should be able to identify and react to situations, which are not pre-defined using the offline rules. Overall, the (near) real-time identification of rules complements the offline rule identification and enhances the performance of the manufacturing enterprise. Consequently, the existing rule identification and validation techniques need to be

 ISBN: 978-1-61208-181-6

transformed to (near) real-time activities based on the actual situations on the shop floor. As a result, the monitoring and control of manufacturing processes becomes more flexible and adaptable.

An event-driven framework has been developed at the Information Systems Institute to minimize the vertical integration gap, and monitor and control manufacturing processes based on complex event processing (CEP). This framework is now been extended to include self-learning monitoring and control mechanisms (i.e., integrate real-time control and real-time rule induction). The remaining part of the contribution is organized as follows. Section II presents research carried out in the area of manufacturing execution systems (MES), CEP, KM, and rule induction. An approach is envisaged in Section III to realize self-learning monitoring and control of manufacturing processes. The implementation of the envisaged approach has been started in an industrial scenario. This scenario is discussed in Section IV. Finally, Section V presents conclusions and outlines future work.

II. STATE-OF-THE-ART

According to VDI 5600, an enterprise can be classified into different enterprise levels [4]. These enterprise levels are (still) inadequately integrated [5], which hinders the establishment of a holistic control of manufacturing processes [1]. Research and development has been carried out to reduce the vertical integration gap between enterprise levels. Software vendors provide MES solutions to bridge the vertical integration gap (e.g., [6][7][8]). However with these MES solutions, major issues still exist with respect to the interface between enterprise levels [5][9]. The exchange of data between enterprise control level and manufacturing execution (i.e., shop floor) is done manually or at most semi-automatically due to inflexible and proprietary interfaces [10]. Hence, standardization activities by several organizations have been performed concerning MES (e.g., [4][11][12]). Latest standardization copes with the definition of logic interfaces for machine and plant control [13].

Event driven architectures (EDA) have been introduced along with MES systems to realize the requirements of real-time monitoring and control of manufacturing processes [10][14]. FORCAM, a MES vendor, uses CEP technology as an innovative approach to monitor, analyze, and control manufacturing processes [15]. The introduction of EDA and CEP engines will assist to separate the control logic (i.e., event processing logic) from the coded application logic. Overall, this will result in an increasing flexibility and adaptability of the monitoring and control systems [16].

Rules are managed in an EP engine as event patterns. Further, the event patterns are formalized using means like event processing language (EPL) statements [10][14][16]. This knowledge is often domain specific and experts are in charge to define proper rules and statements. Knowledge management (KM) can be employed to assist experts to accomplish the aforementioned tasks [17]. For instance, the knowledge discovery in databases (KDD) process can be used to extract control-related knowledge from stored process data [17] and numerous KM tools are available to guide experts with user-friendly interfaces (e.g., [18]). However, KM tools are often utilized offline and consist of several non-trivial activities.

The ability of the control system to rapidly adapt to critical situations during manufacturing process execution is fairly limited (i.e., monitoring and control system has to be adapted manually by modifying the rule base). Classification rule induction is part of machine learning [19][20] and aims to generate a set of classification rules for a given training data set [21]. Direct methods like RIPPER [22] and CN2 [23] derive rules directly from the (process) data. In contrary, indirect methods extract rules by using classification methods like traversing of decision trees. A survey of top-down induction of decision trees classifiers has been presented [24]. Also, parallelizing of classification rule induction has been discussed [25]. The rule induction techniques have been applied in various domains like chemical process control, financial industries, diagnosis of mechanical devices, and classification of celestial objects [26].

The integration of CEP with machine learning (e.g., rule induction) has not been (extensively) explored in literature and industry [27]. A monitoring solution for application, web and database servers has been presented, which is based on the integration of CEP with the machine learning algorithm FRAHST [27][28]. Also, credit card fraud detection based on a combination of CEP with various machine learning techniques (e.g., discriminant analysis, hidden Markov models) has been investigated [29].

III. APPROACH FOR SELF-LEARNING MONITORING AND CONRTOL

An overview of a self-learning monitoring and control system of manufacturing processes is depicted in Figure 1. The central idea of the system is to couple EP with machine learning techniques (i.e., rule induction). Production resources at the shop floor generate process data that denotes quality of products, parameters of resources, and performance of manufacturing processes, among others. This process data is collected by a data collection engine, which implements various industrial communication protocols (e.g., Modbus) [1]. Next, a data aggregation engine aggregates the collected process data with the data from enterprise applications (e.g., order details from ERP system) and builds tracking objects. Each tracking object represents a certain process entity (e.g., order, product) [30].

A tracking object can also be interpreted as a (complex) event by a CEP engine [31]. This CEP engine analyzes the incoming event streams (i.e., integrated process data as tracking objects) and detects (pre-defined) critical situations, and thus, monitors the manufacturing processes. The CEP engine deduces appropriate actions to control the underlying manufacturing processes in case of detection of critical situations. The action can be a combination of (i) displaying alarm messages with assistance of charts and gauges or via communication channels like emails and SMS, (ii) advising operators to modify resource parameters, (iii) manipulating process parameters in the controller of a resource.

Nevertheless, there are some instances where the CEP engine detects critical situations during process execution

 ISBN: 978-1-61208-181-6

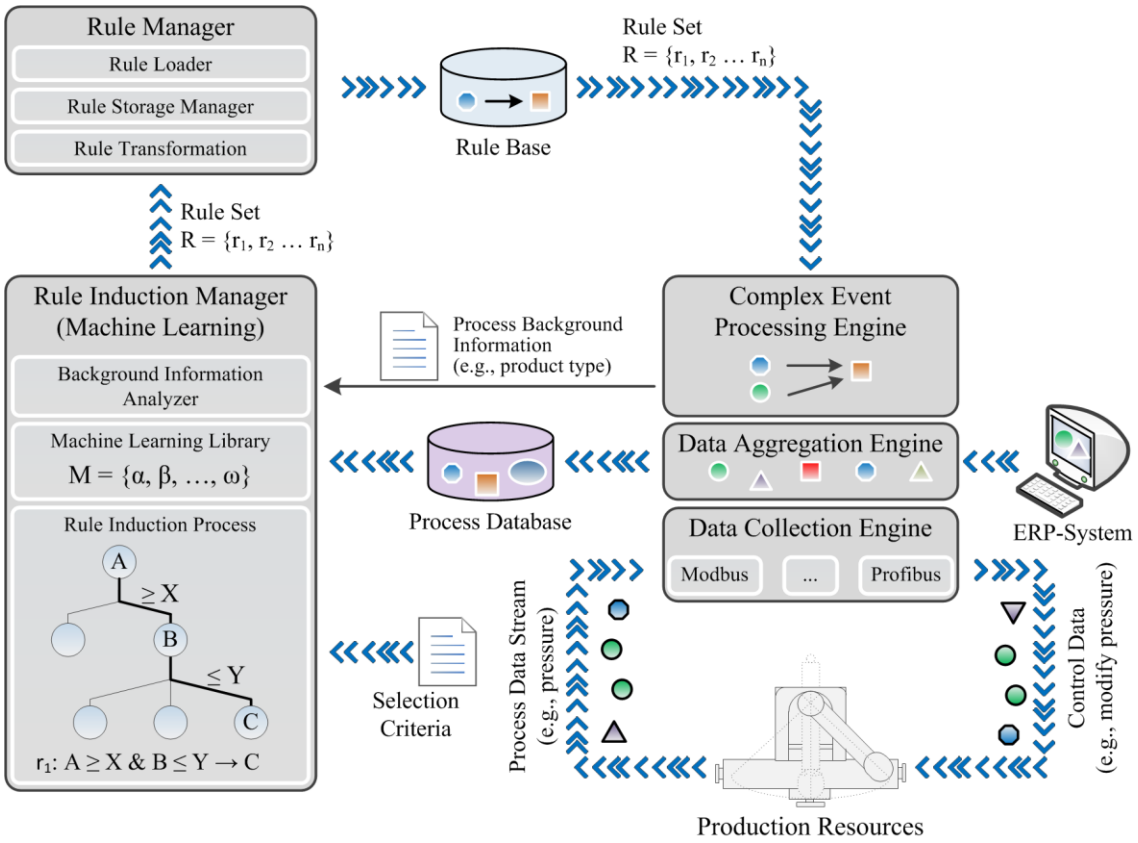


Figure 1. Overview of self-learning monitoring and control system.

(e.g., number of product rejects has exceeded a certain threshold), but cannot provide a proper suggestion to the operators to overcome the identified situations or even manipulate the resource parameters. The manufacturing ramp-up of a novel product can be an example for such an instance. In this situation, the CEP engine initiates a process in the rule induction manager (see Figure 1) and forwards current process background information (e.g., product information) describing the context of the considered manufacturing process. The aim of the rule induction manager process is to refine / improve the rule base / event patterns employed in the CEP engine. The refinement of the rule base is performed by employing machine learning techniques (e.g., decision trees).

The rule induction manager follows a sequence of activities to refine / improve the rule base. First, the process background information (i.e., process context) is analyzed and used to select a limited process data sample from the process database. For instance, the sample size can be restricted to select data for a specific product type and given time range. This step is mandatory as a huge sample size can overstrain the rule induction process (i.e., generation of rules). In addition, parallelization of classification algorithms can be considered to reduce the computation time [25].

Second, a suitable algorithm has to be selected from a rule induction library. A concise overview of rule induction techniques, which are promising candidates related to EP are listed [29]. The current research focuses on (classification) rule induction as rules are transparent and interpretable for domain experts [21]. The selection and parameterization criteria (e.g., rule accuracy, rule coverage) of a certain rule induction algorithm is defined in an XML configuration file of the rule induction manager, and thus, can be suitably modified by domain experts.

Third, rules are generated by employing the selected rule induction algorithm. This can be either performed by a direct method (e.g., RIPPER) or indirect method (e.g., traversing of a decision tree [24]). The derived rules are evaluated against the predefined criteria. Further, the previously evaluated rules can be pruned to obtain more general rules (i.e., cover more instances of the sample data set).

Finally, the induced rules are added to the rule storage (e.g., XML file format) and loaded into the CEP engine as EPL statements. The added rule complements the available rules, which might have been derived offline or online. The default action for a newly added rule is to visualize it as an alarm message because production resources should not be automatically manipulated without operator's awareness. Nevertheless, an operator can modify the generated rules and define suitable actions (e.g., directly manipulate production resources).

IV. INDUSTRIAL CASE STUDY

The aforementioned methodology for a self-learning monitoring and control system elaborated in Section III has been (partly) put into practice in a casting enterprise [32],

 ISBN: 978-1-61208-181-6

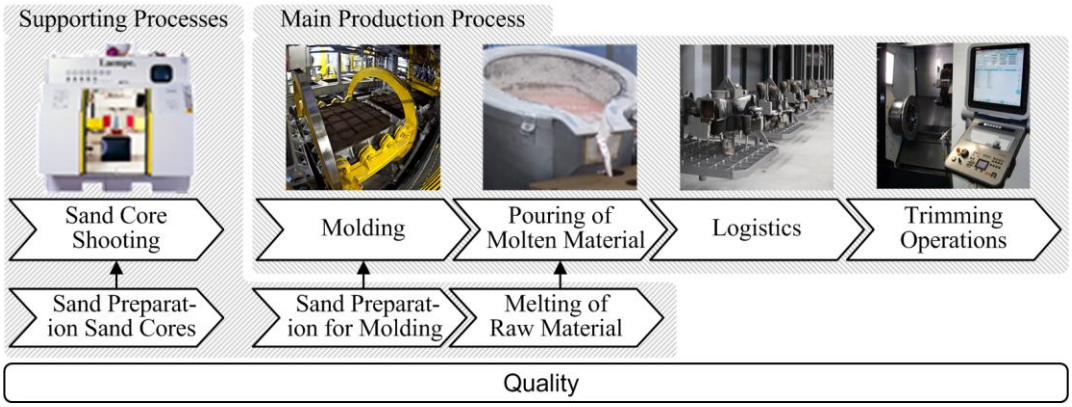


Figure 2. Simplified view of a highly automated casting process.

with the casting process as illustrated in Figure 2. The enterprise in consideration is characterized by a high mix production and low volume production (i.e., from few castings to thousands of castings per order). A highly automated molding machine is employed, which assists to realize the aforesaid characteristics.

This machine simultaneously produces upper and lower molds. The molds are manually inspected for contour and surface damages after a certain number of lower and upper molds have been produced, which is mainly due to (construction) constraints in the manufacturing system. However in case of rejection during inspection, there is a high probability that a certain number of lower molds (here: 30 molds) following the rejected molds would have similar damages.

The aforementioned situation will have a negative influence on the production performance – lower utilization of resources, material wastage and above all a declined commitment to customers. To overcome this situation, a self-learning monitoring and control of manufacturing processes is indispensable. An event-driven framework for enabling EI has been implemented using the Microsoft™ Visual Studio IDE and the .NET framework 4.0.

Real-time process data from the shop floor along with the data from enterprise applications is integrated and stored in an Oracle® 10g database. Further, the integrated data is utilized to create online tracking objects. The integrated data and online tracking objects are forwarded to process visualization clients, which display those using charts and gauges for monitoring purposes. In addition, the online tracking objects are utilized for monitoring and control of manufacturing processes using a state-of-the-art EP engine. Here, EsperTech™ EP engine is employed [33].

The EP engine uses event patterns for the detection of (critical) situations in the process data streams. By default, a rule base has been initialized with rules defined by domain experts. The main goal of the control system is to reduce the number of rejects. If this number increases without any (re-) action of the control system, a machine learning process will be automatically initiated. Classification rule induction methods have to be employed to deduce rules, which can be used to mitigate the aforementioned situation.

V. CONCLUSION AND FUTURE WORK

Manufacturing enterprises are enhancing their monitoring and control of manufacturing processes to sustain their competitive advantages. EI can be utilized to have a holistic view of the enterprise. This EI needs to be exploited to enhance the monitoring and control of manufacturing processes. Further, research has been done to incorporate an EP engine to monitor and control manufacturing processes in (near) real-time. Here, the rules are managed as event patterns and event patterns are used to analyze the incoming process data streams. However, the rule identification and validation process is an offline activity. The existing rules do not adapt whenever there is a change in the processes' situation. Consequently, the offline activity needs to be transformed into a (near) real-time activity.

In the current contribution, an approach has been presented to identify and validate rules using rule induction techniques. On detection of certain pre-defined situations, the CEP engine triggers a sequence of steps in the rule induction manager and at the same time forwards current process background information. This sequence encompasses: (i) selecting suitable process data, which is restricted to the background information, (ii) choosing a suitable rule induction algorithm and defining selection criteria, (iii) identification and evaluation, and generalization of identified rules, and (iv) transformation of the selected rules into event patterns. The presented approach has been (partly) implemented in a casting enterprise, especially with the aim to reduce the rejection of lower molds.

ACKNOWLEDGMENT

Parts of the work presented here have been supported by German Federal Ministry of Economics and Technology (BMWi) as part of "Central Innovation Programme SME" (ZIM) initiative (KF2111502LL0). Also, we are thankful to our industrial partner Ohm & Häner Metallwerk GmbH & Co. KG, Germany for the opportunity to implement and validate the elaborated framework in a casting enterprise.

REFERENCES

[1] M. Grauer, D. Metz, S. Karadgi, W. Schäfer, and J. W. Reichwald, "Towards an IT-Framework for Digital Enterprise

Integration," Proc. 6th Int. Conf. on Digital Enterprise Technology (DET 2009), AISC, vol. 66, Springer, Berlin, Dec. 2009, pp. 1467-1482.

[2] ISO 15704, Requirements for Enterprise Reference Architecture and Methodologies, ISO 15704:2000/Amd 1:2005, 2005.

[3] S. Chakravarthy and R. Adaikkalavan, "Provenenace and Impact of Complex Event Processing (CEP): A Retrospective View," it – Information Technology, vol. 51, no. 5, 2009, pp. 243-249.

[4] VDI 5600, Manufacturing Execution System (MES) - VDI 5600 Part 1, Verein Deutscher Ingenieure (VDI), 2007.

[5] H. Panetto and A. Molina, "Enterprise Integration and Interoperability in Manufacturing Systems: Trends and Issues," Computers in Industry, vol. 59, no. 7, 2008, pp. 641-646.

[6] J. Kletti, Ed., Manufacturing Execution System – MES, Springer, Berlin, 2007.

[7] MPDV, http://www2.mpdv.de/en/ (Accessed: 14.09.2011).

[8] S. Karnouskos, O. Baecker, L. de Souza, and P. Spiess, "Integration of SOA-Ready Networked Embedded Devices in Enterprise Systems via a Cross-Layered Web Service Infrastructure," Proc. 12th IEEE Int. Conf. on Emerging Technology and Factory Automation (ETFA 2007).

[9] B. Saenz de Ugarte, A. Artiba, and R. Pellerin, "Manufacturing Execution System – A Literature Review," Prod. Plan Control, vol. 20, no. 6, Sept. 2009, pp. 525-539.

[10] M. Grauer, S. Karadgi, D. Metz, and W.Schäfer, "An Approach for Real-Time Control of Enterprise Processes in Manufacturing using a Rule-Based System," Proc. Multikonferenz Wirtschaftsinformatik (MKWI 2010), pp. 1511-1522.

[11] VDMA 66412, Manufacturing Execution Systems (MES) - Key Performance, 2009.

[12] IEC 62264, Enterprise-Control System Integration, All Parts.

[13] VDI 5600, Manufacturing Execution System (MES) - VDI 5600 Part 3, Verein Deutscher Ingenieure (VDI), 2011.

[14] S. Karadgi, D. Metz, M. Grauer, and W. Schäfer, "An Event Driven Software Framework for Enabling Enterprise Integration and Control of Enterprise Processes," Proc. 10th Int. Conf. on Intelligent Systems Design and Applications, Cairo, 2010, pp. 24-30.

[15] FORCAM, http://www.forcam.co.uk (Accessed: 14.09.2011).

[16] O. Etzion and P. Niblett, Event Processing in Action. Manning, Greenwich, 2011.

[17] M. Grauer, D. Metz, S. S. Karadgi, and W. Schäfer, "Identification and Assimilation of Knowledge for Real-Time Control of Enterprise Processes in Manufacturing," Proc. 2nd Int'l Conf. on Information, Process and Knowledge Management (eKNOW 2010), Feb. 2010, pp. 13-16.

[18] IBM SPSS Modeler Professional, http://www.spss.com/, (Accessed: 14.09.2011).

[19] D. Michie, D. Spiegelhalter, and C. Taylor, Eds., Machine Learning, Neural and Statistical Classification. Ellis Horwood, Chichester, 1994.

[20] R. Michalski, J. Carbonell, and T. Mitchell, Eds., Machine Learning: An Artificial Intelligence Approach, Volume I . Tioga, Palo Alto, CA, 1983.

[21] P. Flach and N. Lovrac, Rule Induction, http://www.cs.uu.nl/docs/vakken/adm/RuleInduction.pdf, (Accessed: 14.09.2011).

[22] W. Cohen, "Fast Effective Rule Induction," Proc. 12th Int. Conf. on Machine Learning, 1995.

[23] P. Clark and T. Niblett, "The CN2 Induction Algorithm," Machine Learning, vol. 3, no. 4, Springer, 1989, pp. 261-283.

[24] L. Rokach and O. Maimon, "Top-Down Induction of Decision Trees Classifiers – A Survey," IEEE Trans. on Systems, Man, and Cybernetics – Part C: Applications and Reviews, vol. 35, no. 4, 2005, pp. 476-487.

[25] F. Stahl, M. Bramer, and M. Adda, "PMCRI: A Parallel Modular Classification Rule Induction Framework," Lecture Notes in Computer Science, vol. 5632, 2009, pp. 148-162.

[26] P. Langley and H. A. Simon, "Applications of Machine Learning and Rule Induction," Communications of the ACM, vol. 38, 1995, pp. 55-64.

[27] P. H. dos Santos Teixeira, R. G. Clemente, R. A. Kaiser, and D. A. Vieira Jr, "HOLMES: An Event-Driven Solution to Monitor Data Centers Through Continuous Queries and Machine Learning," Proc. 4th ACM Int. Conf. on Distributed Event-Based Systems (DEBS 2010), New York, NY, USA, 2010, pp. 216-221.

[28] P. H. dos Santos Teixeira and R. L. Milidiu, "Data Stream Anomaly Detection through Principal Subspace Tracking," Proc. 2010 ACM Symposium on Applied Computing (SAC 2010), Sierre, Switzerland, 2010, pp. 1609-1616.

[29] A. Widder, R. v. Ammon, P. Schaeffer, and C. Wolff, "Identification of Suspicious, Unknown Event Patterns in an Event Cloud," Proc. ACM Int. Conf. on Distributed Event-Based Systems (DEBS 2007), Toronto, Canada, 2007, pp. 164-170.

[30] M. Grauer, S. Karadgi, and D. Metz, "Enhancement of Transparency and Adaptability by Online Tracking of Enterprise Processes," Proc. 10th Int. Conf. on Wirtschaftsinformatik, Zurich, 2011, pp. 282-291.

[31] M. Grauer, B. Seeger, D. Metz, S. Karadgi, and M. Schneider, "About Adopting Event Processing in Manufacturing," Lecture Notes in Computer Science (LNCS) 6569, M. Cezon and Y. Wolfsthal, Eds. Springer, Berlin, 2011, pp. 180-187.

[32] M. Franken, "Welcome to 21st Century," Giesserei, vol. 97, 2010, pp. 90-99 (in German).

[33] EsperTech, http://www.espertech.com/ (Accessed: 14.09.2011).

 ISBN: 978-1-61208-181-6

An Overview of Norwegian Linked Open Data

~ Applications in Regional Development and Environmentally Friendly Behavior ~

Dumitru Roman
SINTEF
Oslo, Norway
Dumitru.Roman@sintef.no

David Norheim
Computas AS
Lysaker, Norway
David.Norheim@computas.com

Abstract— **With Norway being one of the few countries outside of the English speaking world with a clear governmental strategy and commitment to open data, combined with one of the highest Internet penetration and mobile access in Europe, it offers interesting opportunities for becoming a great testbed for consuming Linked Open Data (LOD). With this paper we aim at presenting potential applications consuming Norwegian LOD and showing practical benefits of aggregating open data in highly sensitive domains for governments and the general public such as regional development and environmentally friendly behaviour. At the same time, this paper will serve as an overview of the Norwegian LOD as of mid 2011. The proposed applications will not only aim at demonstrating the benefits of the current Norwegian LOD, but will also make contributions to the improvement and extension of the existing data sets.**

Keywords-Norwegian LOD, data sets, applications

I. INTRODUCTION

Norway is one of a handful of countries outside of the English speaking world with a clear commitment to open data. Being one of the first countries to implement the PSI-directive as a law in January 2009 [1], each new governmental project is today required to address publication of the data it creates or processes. Simultaneously the community focusing on linked open data is ever increasing, and major governmental agencies are involved in making their data available as Linked Open Data (LOD). However, as we write 2011, few if any, applications are using LOD as their data source. We expect this to change soon.

We believe that the interest in LOD, combined with the highest Internet penetration in Europe after Iceland, and second only to Sweden on mobile internet access in Europe, Norway offers interesting opportunities for becoming a great testbed for consuming LOD data. With this paper we aim at giving an overview of the current status of the Norwegian LOD, and propose applications consuming Norwegian LOD in highly sensitive domains for governments and the general public such as regional development and environmentally friendly behaviour. The proposed applications will not only aim at demonstrating the benefits of the current Norwegian LOD, but will also make contributions to the improvement and extension of the existing data sets.

In the following we provide an overview of two potential applications consuming Norwegian LOD (Section II), and then present the current LOD in Norway, with a focus on its relevance to the proposed applications (Section III). Section III summarizes this paper.

II. APPLICATIONS FOR NORWGIAN LOD

A. Regional development with LOD

Why? Regions are often faced with challenges such as lack of competitiveness and need for regeneration to various political challenges. The long-term trends and effects of development schemas are however often not well understood and readily available to the public. Local news agencies and the general public are often served numbers without being able to look behind the scene. There is a clear need in this domain for support in creating visualization tools over statistical data or aggregated data to follow the long term effects of regional development.

Who are the stakeholders? Data-journalists in Norway, both from the broadcasters and news agencies, have requested a tool allowing ad-hoc combinations of datasets in their regional foot print. A tool based on linked data will increase its value as new datasets become available. Data visualization and investigation support should also be available for public use, allowing citizens to drill into public data.

What and how? Inspired by Hans Roslings famous presentation tool Gapminder [2], we propose to build an application focusing on innovations and development in various sectors in regions and municipalities. Central Norwegian public sector data owners have opened their data as linked open data during 2010. The Brønnøysund Registry Center [1] has opened the national organization registry, Enhetsregisteret [3]. Others, notably the Norwegian Research Council, [2] have already used the dereferencable URIs from Enhetsregisteret while publishing their own data sets. This case study will require the need to improve the existing company registry service (exporting also NACE codes), and to include geographical information (linking to Kommuneregisteret,[3] the details of municipalities and counties). In addition it requires the use of visualization tools. The successful implementation of

[1] http://www.brreg.no/english
[2] http://www.forskningsradet.no/en
[3] http://www.kommunenokkelen.no/adresse/side2.do?side=kommuneregisteret

 ISBN: 978-1-61208-181-6

this application is dependent on the quality of the data and links in the existing Norwegian LOD, and on useful and flexible analytics and visualization of the linked data which will be developed as part of this case study.

B. *Assisting Environmentally Friendly Behaviour with LOD*

Why? The current Norwegian government has made a point that environmentally friendly behaviour should pay off. However, behaving environmentally friendly is not an easy task even if financial incentives are in place. For example, often the information about public transportation and availability of city bikes is not as easily available when a trip decision needs to be taken. In such situations, there is a need for applications that combine linked open data in the environmental domain as a decision making tool.

Who are the stakeholders? Citizens interested in applications assisting them in environmentally friendly behaviour.

What and how? In the private sector a number of applications and mobile apps have been created as a result of open data initiatives. In Norway this includes applications ranging from real-time public transportation information [4], snow and ski conditions [5], the presence of electric cars charging stations [6], real time status for city bike stands [7], weather forecasts [8], and many more. Common to all this is that they use only *one* source of open information. The proposed application in the environmental domain will combine the use of transportation (e.g. public transportation, electric cars parking lots, bikes stands) to events (e.g. concerts, art galleries) while including other decision relevant real-time information (e.g. forecast, traffic messages). The application will be made as a mobile app. The overall goal here is to invoke the citizen's own interest in environmentally friendly behaviour. The successful implementation of this application is dependent on extending the existing Norwegian LOD with new data sets and ensuring the quality of the new data sets and their proper linking to the existing Norwegian LOD, as well as on the usability of such an environmentally friendly behaviour application which will be developed as part of this case study.

III. NORWEGIAN LOD

On January 1st 2009 the EU PSI Directive was implemented in Norwegian law. As a direct consequence of this, the Ministry of Governmental Reform and Administration in 2010 announced the development of data.norge.no (currently in alpha mode), the equivalent of data.gov and data.gov.uk, to be launched during spring 2011. While the portal will be filled with PSI data sets, other data sets will be included such as those provided two central Norwegian LOD projects, Sesam4[4] and Semicolon II[5], which opened data sets and converted them to RDF with links in-between and to other sources. Of special interest was the national registry for company information, Enhetsregisteret, and the archive for research funding, Prosjektdatabasen, from the Norwegian Research Council.

Figure 1 provides an overview of the existing data sets in the Norwegian LOD and their relationships. The blue-coloured data sets are related to the regional development case study and the green-coloured data sets are related to the environmentally friendly behaviour. The solid arrows between the data sets represent the already existing links between the data sets, whereas the dotted arrows represent the logical connections for which we plan to create the necessary links in the project (as far as the relevant data sets are concerned).

A brief description of the relevant data sets we are considering for the two applications is given below:

- All legal entities are registered in the **Central Coordinating Register for Legal Entities (Enhetsregisteret)** at Brønnøysund Registry Centre. URIs and a RESTful service returning RDF has been created to look up based on organizational number. The service needs to be made searchable by including a RESTful service based lookups on NACE codes and regional codes.
- The Norwegian Association of Local and Regional Authorities (KS) and Statistics Norway maintain a **registry of municipalities and counties (Kommunekatalogen)**. The information translated to RDF as a proof of concepts by the research project Semicolon II, but URIs haven't been made RESTful and outgoing links to e.g. dbPedia has not yet been created.
- The Norwegian Research Council maintains a **catalog of funded projects (Prosjektarkiv)**. This relates instruments and projects to organisations. This catalog has been translated to RDF and outgoing URIs have been created to the Enhetsregisteret. The catalog is hosted on a triplestore supplied by Computas through the SESAM4 research project.
- All the parties are, according to the Political Parties Act, obliged to report their income to a central register. The Ministry of Government Administration, Reform and Church Affairs, maintains a **central register of parties and their income (Partifinansiering.no)**. This data is available in RDF / SPARQL, on Computas triple store and have been assigned URIs, but has not yet been linked to the Kommunekatalogen.
- The Ministry of Local Government and Regional Development have made **election results** from elections available in structured form. These data should easily be translated to RDF with outgoing links to the party register mentioned above and then made available in a triple store (SPARQL).

[4] http://sesam4.net/
[5] http://www.semicolon.no/

 ISBN: 978-1-61208-181-6

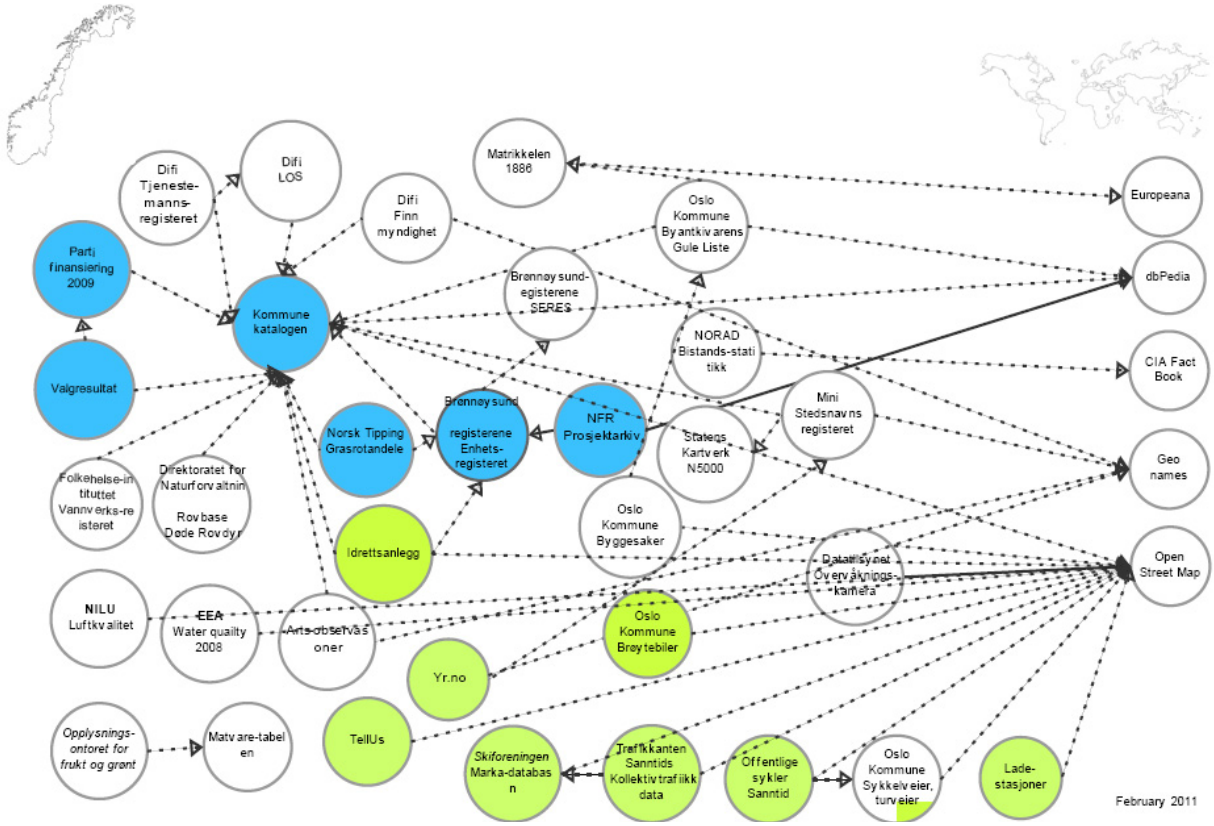


Figure 1. Norwegian LOD.

- Norsk Tipping is a wholly state-owned company under the jurisdiction of the Ministry of Culture with a special role in regulating gaming in Norway. Gaming funds benefit sporting and cultural activities as well as voluntary humanitarian organisations all over the country through the Grassrootshare (Grasrotandelen). Grasrotandelen is a **registry showing organisations that receive funding**. The registry has been transalated to RDF but without the links to the organisations involved (in Enhetsregisteret).

- Trafikanten's main purpose is to provide the public with **up to date travel information for the public transportation system** in and around the counties Oslo and Akershus. The data is about all public transportation, the current location and delays of route. Currently the data is not available to the public as raw data (only through their own app), but there are currently work going on in Oslo Municipality to make it available as RDF.

- Yr.no is the joint online **weather** service from the Norwegian Meteorological Institute (met.no) and the Norwegian Broadcasting Corporation (NRK). They have a XML service available for 7 million places in Norway for free. Work is going on in the Semicolon II project to make this available as RDF RESTful services.

- Markadatabasen is maintained by the Norwegian Ski Association. It's a public source that provides **routes and real time status of preparation of ski slopes**. The data is available in XML. It is also linked to Open Street Map. The data needs to be translated to RDF and made available as a RESTful and SPARQL endpoint. The data will also be linked to public transport routes from Trafikanten.

- Oslo is equipped with around 3000 city-bikes available for the public. A **real-time status of the bike stands** (where bike stands are and how many are available) is developed. ClearChannel is maintaining this on behalf of Oslo Kommune. The data is made available as a XML service. The service needs to be wrapped as RDF and interlinked with bike tracks and tour tracks.

- **Bike tracks and touring tracks** in Oslo is a dataset that is made available from Oslo Kommune. This should be connected to the city-bike stands and made available as a SPARQL/RDF endpoint.

- All national **electric cars charging stations** for electric-cars are made publicly available as a dataset and translated to RDF, and made available by the research project SESAM4 on a triple store. It is now being extended to include real-time availability. This real-time service will have to be interfaced to the charing station locations.

 ISBN: 978-1-61208-181-6

- **Snow ploughing real-time status** is another service that is made available by the Oslo Kommune in near future. This data set could relate to traffic information making it more sensible to take public transportation. The dataset should therefore link to public transportation and made available in RDF, and also connected to the weather information.
- Tellus is a national dataset of **accommodations, attractions and events** produced by the same organisation. It has been made available as a SPARQL endpoint with RESTful RDF lookup by the Sesam4 project. It should be linked to Open Street map and possibly to public transportation.
- The Ministry of Culture maintains a dataset of **all sports arenas** in Norway. This dataset has been converted to RDF and made available as SPARQL endpoint with RESTful lookup in RDF by the Sesam4 project. This should be connected to the organisation in the Enhetsregisteret.

Table 1 below provides information of the above mentioned data sets regarding the owners, the current formats they are available in, where they are made available, the estimated size, the quality of data, and the case study where they will be applied.

Table 1. Relevant Norwegian LOD data sets for the proposed applications.

Name	Owner	Format	Hosting	Estimated # of triples	Quality (stars)	Application
Enhetregisteret	Brønnnøysundregisterene	Restful RDF Web service	Brønnøysundregisterene	> 4.500.000	5	App #1
Kommunekatalogen	KS	XML	Univ of Oslo / Semicolon II	Ca 2.000	3	App #1
NFR prosjektarkiv	Norwegian Research Council	RDF	Computas/ Sesam4	Ca 200.000	5	App #1
Valgresultat 2005	Government	RDF	Computas/ Semicolon II	Ca 100.000	4	App #1
Partifinansiering 2009	Government	RDF	Computas/ Semicolon II	Ca 100.000	4	App #1
Grasrotandelen	Norsk Tipping	RDF	Computas/ Semicolon II	Ca 70.000	4	App #1
Trafikkanten sanntid	Oslo Kommune	XML, JSON	Oslo Kommune or Computas	Ca 50.000	3	App #2
Yr.no	Met. Inst.	XML	Univ of Oslo / Semicolon II	Ca 700.000.000	3	App #2
Markadatabasen sanntid	Skiforeningen	XML	Oslo Kommune or Computas	Ca 100.000	3	App #2
Offentlige sykler sanntid	ClearChannel	XML	Oslo Kommune or Computas	Ca 10.000	3	App #2
Sykkelveier, turveier	Oslo Kommune	XML	Oslo Kommune or Computas	Ca 10.000	3	App #2
Ladestasjoner (Sanntid)	Ladestasjoner.no	RDF	Computas/ Semicolon II	Ca 1.100	3	App #2
Brøytebiler Sanntid	Oslo Kommune	XML	Oslo Kommune or Computas	Ca 10.000	3	App #2
Tellus	Tellus	RDF	Computas/Sesam4	Ca 600.000	3	App #2
Idrettsanlegg	Ministry of Culture	RDF	Computas/Sesam4	Ca 1.000.000	3	App #2

IV. SUMMARY AND OUTLOOK

This paper gave an overview of the Norwegian LOD as of mid 2011, and proposed two applications consuming the Norwegian LOD. The aim of the two applications is to enhance transparency in making factual information available to the public when it comes to regional development, and help individuals to act more environmentally friendly. We outlined the data sources available for building these two applications. The two applications rely on local and national data sets, but are highly generic and could be applicable for other countries as well.

Whereas the data is available, improving its quality is of crucial importance for a successful realization of the proposed applications. In particular, the quality of links between the relevant data sets needs to be ensured. The links should be created with quality information based on the method being used, e.g. lookups in other registries, ontology reasoning, etc, and be put into a separate graph holding this provenance information. In addition, especially for the second application, available open data sets need to be published as RDF in a triple store and an infrastructure for the triple store and RESTful services needs to be available.

ACKNOWLEDGMENT

This work is partially funded by the nationally funded Semicolon II project, and the EU funded projects PlanetData and ENVISION.

REFERENCES

[1] Implementation of the PSI-directive: http://www.lovdata.no/all/hl-20060519-016.html#9 (In Norwegian).

[2] Gapminder: http://www.gapminder.org.

[3] Enhetsregisteret: http://www.brreg.no/registrene/enhet.

[4] Trafikanten sanntid: http://itunes.apple.com/no/app/trafikanten-sanntid/id299318111?mt=8.

[5] iMarka: http://itunes.apple.com/no/app/imarka/id344496606?mt=8.

[6] LadeNå!: http://itunes.apple.com/no/app/id329169428?mt=8.

[7] Oslo Bysykkel: http://www.oslobysykkel.no/, Accessed April 2011.

[8] Yr.no: http://itunes.apple.com/us/app/yr-no/id300709016?mt=8.

 ISBN: 978-1-61208-181-6

Critical Dimension in Data Mining

Divya Suryakumar, Andrew H. Sung
Department of Computer Science and Engineering
New Mexico Institute of Mining and Technology
Socorro, New Mexico 87801, USA
divya|sung @cs.nmt.edu

Qingzhong Liu
Department of Computer Science
Sam Houston State University
Huntsville, Texas 77341, USA
liu@shsu.edu

Abstract - **Data mining is an increasingly important means of knowledge acquisition for many applications in diverse fields such as biology, medicine, management, engineering, etc. When tackling a large-scale problem that involves a multitude of potentially relevant factors but lacking a precise formulation or mathematical characterization to allow formal approaches to solution, the available data collected for the application can often be mined to extract knowledge about the problem. Feature ranking and selection, thereby, are immediate issues to consider when one prepares to perform data mining, and the literature contains numerous theoretical and empirical methods of feature selection for a variety of problems. This work in progress paper concerns the related question of critical dimension, i.e., for a specific data mining task, does there exist a minimum number (of features) which is required for a specific learning machine to achieve satisfactory performance? As a first step in addressing this question, a simple ad-hoc method is employed for experiment and it is shown that the phenomenon of critical dimension indeed exists for several of the datasets studied. The implications are that each of these datasets contains irrelevant features or input attributes, which can be eliminated to achieve higher accuracy in model building using learning machines.**

Keywords-feature selection; critical dimension; machine learning.

I. INTRODUCTION

Data mining is aimed at extracting useful information or knowledge from datasets; to achieve this goal, feature selection is often necessary to eliminate lesser or insignificant features in order to reduce the size of the dataset and to facilitate model building (e.g., using learning machines) for knowledge extraction. Many methods have been proposed for feature selection [1]. The interesting fact about extracted features are that sometimes not all extracted features are individually useful; however, correlation of features itself an intriguing question.

We may use learning machines to find feature correlation or to discover important or relevant features. Some theoretically optimal criteria could become practically intractable [2]. The ultimate, guaranteed optimal feature selection method requires exhaustive analysis of all possible subsets of features; this is infeasible for datasets with a large number of features; so, the next best goal is to find a satisfactory set of subsets. Feature selection is usually done in two different ways, namely subset selection or entropy-based selection and feature ranking. Feature ranking uses ranking algorithms which scores all features using certain metrics and ranks them accordingly [3]. A subset selection method uses an algorithm to find a best possible subset in arbitrary time. Here, the term best possible subset refers to the best subset found among satisfactory set of subsets [4].

II. FEATURE RANKING

The main objective of feature selection is to improve the prediction performance or accuracy, to provide faster and cost-effective predictors and understand the correlation among data [5]. For our experiments, we use both feature selection and subset selection.

A supervised 'Chi-squared Ranking Filter' [6] and a supervised 'Support Vector Machines (SVM) feature evaluator' [7] method are used for ranking features. A 'Ranker' search method ranks attributes according to their relevance and individual evaluations. Using Ranker we can set the threshold to reduce the attribute set to consider or also specify the set of attributes to ignore; hence it is comfortable for our experiments to eliminate some unwanted features. The Chi-squared Ranking Filter evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to the class. It is a statistical test to find the independence of two events for goodness of fit of an observed distribution to a theoretical one whose value is in zero to infinity range and cannot be negative. SVM feature evaluator evaluates the worth of an attribute by using an SVM classifier. Attributes are ranked by the square of the weight assigned by the SVM feature evaluator. Attribute selection for multiclass problems is handled by ranking attributes for each class separately using a one to all method and then dealing from the top of each pile to give a final ranking.

To find the best feature subset, we use supervised CFS Subset Evaluator method and a greedy stepwise search algorithm. The algorithm evaluates the worthiness of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low inter-correlation are preferred [8][9].

The two feature selection methods discussed above are the most widely used methods but there could always be that one subset which is the best feature subset or the correlation among a certain low ranked features could increase the

 ISBN: 978-1-61208-181-6

performance. Hence, in this paper, we show results of a method called critical dimension which can provide us the minimum number of features that are required for a learning machine to perform accurately.

III. CRITICAL DIMENSION

The *critical dimension* of a dataset is the minimum number of features required for a learning machine to perform prediction or classification with high accuracy. As such, it is an informal concept and empirical methods are called for to determine the critical dimension. Thus critical dimension of a dataset can be defined as that number (of features) where the performance of a specific learning machine would begin to drop significantly, and would not rise again when smaller number of features is used.

Specifically, it is postulated that for a dataset there possibly exists a critical dimension μ which is a unique number for a specific machine learning and feature ranking combination. More clearly, let $A = \{a_1, a_2, \ldots, a_n\}$ be the feature set where $a_1, a_2, \ldots, a_n$ are listed in order of decreasing importance as determined by some feature ranking algorithm. Let $A_m \subseteq A$ contains the m most important features, i.e., $A_m = \{a_1, a_2, \ldots, a_m\}$ where $m \leq n$. For a learning machine *M* and a feature ranking method *R*, we call μ ($\mu \leq n$) the critical dimension of [*M*, *R*], if whenever *M* uses feature set A_k with $k \geq \mu$ the performance of *M* is ≥ T, where T represents a performance threshold deemed satisfactory; and whenever *M* uses less than μ features its performance drops below T; further, *M*'s performance from μ to μ–1 features decreases significantly.

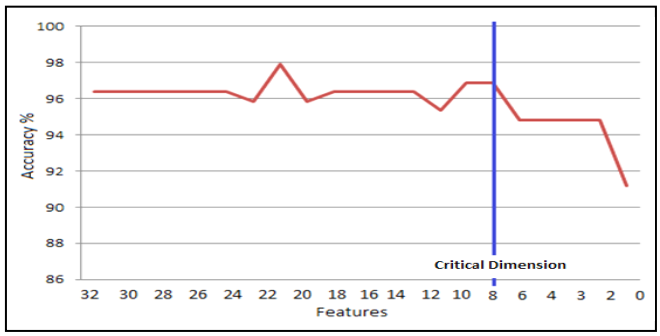


Figure 1. Showing the critical dimension at feature size 8

The graph in Figure 1 shows that there exists a μ at 8 features in the Wisconsin breast cancer dataset [10] dataset.

AdaBoost was used to classify this dataset. The graph is plotted with the number of features on the x-axis and prediction accuracy on the y-axis. From the graph we can see that the performance decreases if we choose lesser features than μ and the performance never rises above the measure at μ.

The first step in find μ is to rank all features using ranking algorithms. In this experiment we used *Chi Squared Attribute Evaluator* as the attribute evaluator and Ranker as the search method for feature ranking and a SVM subset evaluator for subset feature selection. Once the datasets are ranked the prediction accuracy is calculated. In the following iterations prediction accuracy is calculated by removing one least important feature each time till and beyond the critical dimension point. The results are studied and the point at which the performance curve as shown in Figure 1 drops drastically and never rises above that point is defined as the unique μ for that dataset.

Utilizing results from experiments carried out earlier, we can say that μ exist in most datasets and that this μ is a unique number pertaining to that dataset for that particular or specific learning machine classifier and ranking combination. Results using similar classifiers by other experimenters are in the UCI database.

The table below shows the results of experiments performed previously on six different datasets from the UCI repository [11] which, either has an obvious critical dimension (O), or no obvious critical dimension (N/O), as shown in the last column of the table. The classifiers used for classifying the datasets are also shown. The initial condition is when all ranked features or the best subset features are analyzed. For some of datasets, all features are feature ranked and then a learning machine classifier is used to find the accuracy and for others the best feature subset is found and classification accuracy is found using a learning machine classifier. Experiments were performed to find μ by removing one least important feature at the beginning of iteration and calculating the performance accuracy at the end of each iteration. In the table below, the accuracy at μ and accuracy during the first iteration are shown. The classifiers used for each dataset are also tabulated. For the Wisconsin breast cancer dataset (WBDC) two different classifiers were used to experiment. It can also be seen that the critical dimension is unique to that (dataset, machine learning algorithm and ranking) combination.

TABLE I. RESULTS OF BIO-MEDICAL DATASETS

SN	Name	Initial condition		At critical dimension		Classifier	Type
		# of features	*Accuracy %*	*# of features*	*Accuracy %*		
1	WBCD	31	96.3731	8	96.8912	Ada Boost	O
2	Hypothoroid	25	97.3953	18	95.2483	SMO	O
3	SPECT Heart	22	74.1573	3	72.6592	Attribute selected	O
4	SPECTF Heart	44	98.001	10	87.9121	Bagging	O
5	Lung Cancer	56	63.6364	24	63.6364	Multi Boost	O
6	WBCD	31	96.3731	6	96.8549	Multilayer Perceptron	N/O
7	Parkinsons Disease	23	96.9697	5	100	Ada Boost	N/O

a. The dataset used for this experiment are from the UCI repository.

 ISBN: 978-1-61208-181-6

Critical dimension is an innovative and cost effective method to reduce the problems involved in feature selection as it is almost always impossible to find the best possible feature subset possible. The main idea is finding the minimum set of features necessary for the successful development of learning machine classifiers for a given dataset. The results from the table demonstrate that this is indeed the case for the several bioinformatics dataset studied.

We can see from the results presented that there exists a unique critical dimension in some datasets which, when found can reduce the feature dimension, without compensating in performance. The accuracy of performance with all the features and at the critical point for all dataset in Table 1 shows that there is not much difference in the performance.

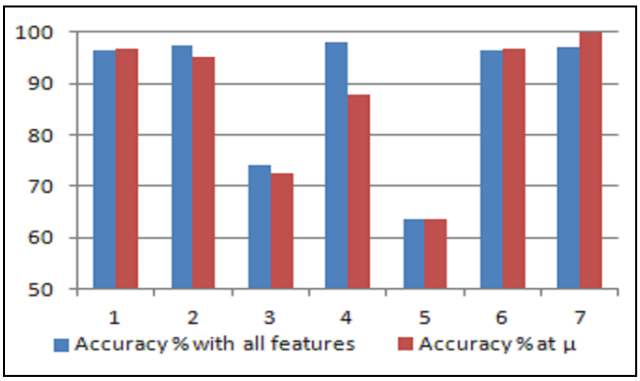


Figure 2. x-axis representing the datasets as numbered in Table 1 and the y-axis are the accuracies

A mushroom dataset was created by us with 127 features. There are two types of genes recorded in this dataset, *Lentinus Fr. and Marasmius Fr.* and 40 samples of each. The 127 features are the mushrooms habitat, details of the macroscopic pileus, macroscopic gills, macroscopic stipe, microscopic context, microscopic basidispore, microscopic basidia, microscopic cystidia, microscopic trama and microscopic Pileipellis. The dataset contains subgenus of both *Lentinus Fr. and Marasmius Fr. Lentinus Fr. and Marasmius Fr.* are the broader classification. The dataset contains some missing values or gaps.

This paper concludes the results of a new method to identify mushroom gene using machine learning methods. Different types of mushrooms are used as an extract to cure certain cancers and hence it is highly important to classify them [12]. In this experiment, we are trying to identify the species into the broader class classification. For example, the *Lentinus Fr.* has subgenus type such as *Lentinus cladopus, Lentinus squarrosulus, Lentinus cyathiformis* etc. which are classified as *Lentinus Fr.* Similarly the subgenus of *Marasmius Fr.* are grouped into type *Marasmius Fr.* Machine learning methods were used to identify the types. This is a binary class classification.

The dataset contains a total of 127 features and 80 samples. The datasets for the experimentation was divided into testing and training sets. The split is 66% for training and the rest for testing. The performance measure was the prediction accuracy of the test set. The mushroom dataset was classified using different classifiers, namely Rule based classifier ZeroR, classifier, SMO, AdaBoost and ADTree. We can see that the rule-based classifier accuracy was poor and SMO and ADTree showed 100% accurate results.

TABLE II. RESULTS OF DIFFERENT LEARNING MACHINE CLASSIFIER

Method	Accuracy%
ZeroR	40.7407
AdaBoost	96.2963
SMO	100
ADTree	100

A ranking algorithm was then used to rank the dataset. The ranking method used was CfsSubsetEval and the selection was made using greedy stepwise algorithm. The output was a feature set of 20 features. The feature numbers of the best feature subset was {3,7,14,22,31,36,58,68,72, 73,74,75,78,84,93,113,121,122,125,127}. Using this best feature subset and SMO classifier the results obtained are shown below.

TABLE III. SMO RESULTS OF BEST FEATURE SUBSET

Method	Accuracy%	Confusion matrix
SMO (using best feature subset)	100	a b 0 16 | a = Lentinus Fr. 11 0 | b = Marasmius Fr.

Now, using Ranker algorithm and ChiSquareAttributeEval method, all 127 features were ranked, for example, the 84th feature was ranked the highest or most important feature and 67th feature was ranked as the least important feature. We then use only the top twenty features ranked by the Ranker and run our learning machine classifier. The dataset was split into training (66%) and testing dataset (34%). The second line shows the output of SMO classifier using the top 20 features.

TABLE IV. RESULTS OF MUSHROOM DATASET

Feature	TP rate	FP Rate	F Measure	ROC Area	Mean abs error	Relative abs error	Accuracy%
31	0.96	0.04	0.964	0.994	0.045	9.758	96.373
30	0.96	0.04	0.964	0.994	0.045	9.758	96.373
11	0.96	0.04	0.964	0.99	0.055	11.903	96.373
10	0.95	0.55	0.953	0.988	0.058	12.416	95.336
9	0.97	0.03	0.969	0.993	0.0562	11.974	96.891
8	**0.97**	**0.03**	**0.969**	**0.993**	**0.0562**	**11.974**	**96.891**
7	0.95	0.07	0.948	0.993	0.0591	12.591	94.818
6	0.95	0.07	0.948	0.993	0.0591	12.591	94.818
5	0.95	0.07	0.948	0.993	0.0607	12.941	94.818
4	0.96	0.03	0.964	0.992	0.0581	12.384	94.818

The critical dimension was found for the mushroom dataset. We can see from the table above that a critical dimension exists and is 7 features. We can see that when the experiment was run using top 6 features the accuracy drops

 ISBN: 978-1-61208-181-6

to 96.2963%. Hence, a critical number 7 can be assigned to this mushroom dataset using SMO.

The above experiments show that a 100% accurate perditions result was obtained by means of a SMO classifier using the best feature subset and also using the same number of features as in the best feature subset, i.e., top twenty features. The dataset was analyzed to find the critical dimension and a feature set containing the top 7 features namely, Microscopic Context type (homoiomerous or heteromerous), presence of Annulus or partial veil, Macroscopic Stipe Color, Microscopic Trama breath, Microscopic Cystidia Cheilocystidia shape, Macroscopic Pileus Shape, and Macroscopic Stipe consistency was found to be the critical dimension of the mushroom dataset. The results of this paper are a breakthrough in mushroom identification for broader genus identification.

The graph showing the critical dimension of the mushroom dataset is shown below. From the plot and the table, we observe that this dataset possesses an obvious critical dimension.

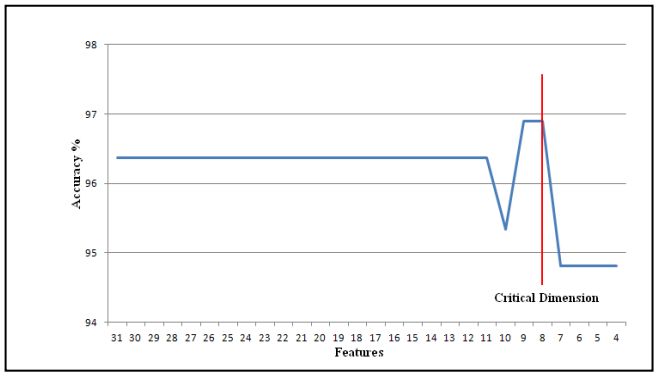


Figure3. Mushroom dataset showing μ = 8

CONCLUSION AND FUTURE WORK

As we continue to explore the concept of critical dimension and seek to develop a more formal framework, we are also trying to study and verify the ramifications of this phenomenon. Clearly, a dataset that exhibits an obvious critical dimension indicates that it contains irrelevant features which can be eliminated, or that the dataset itself is not large enough or sufficiently representative of the problem's whole input space to allow the construction of accurate models using learning-machine-based approaches (i.e., the inclusion of more data points may make the critical dimension disappear). Experiments are also being carried out to study critical dimensions in relation to different learning machines and feature ranking methods, since it appears that the critical dimension of a dataset is dependent on both the adopted learning machine and the adopted feature ranking/selection method for mining the data. It is believed that this research complements the research on feature ranking and selection in several aspects by addressing the question of how many features are essential in building, e.g., a learning machine classifier that delivers acceptable performance. Also, the existence of a critical dimension for a dataset indicates a measure of poor data quality and points to the opportunity of dimension reduction by eliminating useless or irrelevant features.

We are creating a much larger dataset for the mushroom study to perform experiments on multiclass classification and to see if the results are as expected or as good as the binary classification. New dataset will be tested using the top 7 features given by experiments performed in this dataset. Sub genus identification and classification using data mining is the next step after multiclass classification experiments are carried out.

ACKNOWLEDGMENT

Support for this work received from ICASA (Institute for Complex Additive Systems Analysis) of New Mexico Tech and the National Institute of Justice, U.S. Department of Justice (Award No. 2010-DN-BX-K223) and the mushroom dataset provided by CAS in Botany department, University of Madras, Guindy campus, Chennai, India are gratefully acknowledged.

REFERENCES

[1] Dy, J. G. and Brodley, C. E., Interactive visualization and feature selection for unsupervised data, Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 360–364, 2000

[2] Almuallim, H. and Dietterich, T. G., Learning with many irrelevant features, Ninth National Conference on Artificial Intelligence, pp. 547-552, 1991

[3] Guyon, I. and Elisseeff, A., An Introduction to Variable and Feature Selection, Journal of Machine Learning Research 3, pp. 1157-1182, 2003

[4] Dy, J.G., and Brodley, C.E., Feature Subset Selection and Order Identification for Unsupervised Learning, Seventeenth International Conference on Machine Learning, pp. 247-254, 2001

[5] Hong, Z.Q. and Yang, J.Y., Optimal Discriminant Plane for a Small Number of Samples and Design Method of Classifier on the Plane, Pattern Recognition, Vol. 24, pp. 317-324, 1991

[6] Mesleh, A. M. A., Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System, Journal of Computer Science, pp. 430-435, 2007

[7] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V., Gene selection for cancer classification using support vector machines, Machine Learning, pp. 389-422, 2001

[8] Hall, M. A., Correlation-based Feature Subset Selection for Machine Learning, Hamilton, New Zealand, 1998

[9] Geng, X., Liu, T. Y., Qin, T., and Li, H., Feature Selection for Ranking, SIGIR, 2007

[10] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets <retrieved: March, 2010>

[11] Wolberg, W.H., Street, W.N., and Mangasarian, O.L., Machine learning techniques to diagnose breast cancer from fineneedle aspirates, Cancer Letters, Vol. 77, pp. 163-171, 1994

[12] Borchers, A.T., Stern, J.S., Hackman, R.M., Keen, C.L., and Gershwin, M.E., Mushrooms, tumors, and immunity, Exp Biol Med, Vol. 221, pp. 281-293, 1999

 ISBN: 978-1-61208-181-6

Context-aware Recommendation of Visualization Components

Martin Voigt*, Stefan Pietschmann*, Lars Grammel†, and Klaus Meißner*
*Technische Universität Dresden, Chair of Multimedia Technology,
Dresden, Germany
Email: martin.voigt, stefan.pietschmann, klaus.meissner@tu-dresden.de
†University of Victoria, Computer Human Interaction and Software Engineering Lab
Victoria, British Columbia, Canada
Email: lars.grammel@gmail.com

***Abstract*—Although many valuable visualizations have been developed to gain insights from large data sets, selecting an appropriate visualization for a specific data set and goal remains challenging for non-experts. In this paper, we propose a novel approach for knowledge-assisted, context-aware visualization recommendation. Both semantic web data and visualization components are annotated with formalized visualization knowledge from an ontology. We present a recommendation algorithm that leverages those annotations to provide visualization components that support the users' data and task. We successfully proved the practicability of our approach by integrating it into two research prototypes.**

***Keywords*-recommendation, visualization, ontology, mashup**

I. INTRODUCTION

Visualization is a powerful way of gaining insight into large data sets. Therefore, a myriad of visualizations have been developed in recent decades. To bridge the gap between data and an appropriate visual representation, models like the visualization pipeline [1] have been developed and implemented in numerous tools. As one part of this process, the mapping of data to a graphic representation is critical because only small subsets of existing visualization techniques are expressive and effective for the selected data in a specific context. Generally, domain-specific data can be visualized either using tools which were developed specifically for that domain and use case, or using generic visualization systems. The development of the former requires extensive knowledge by visualization and domain experts, and is usually costly and time-consuming. Thus, in many cases generic visualization tools are preferable, because they are quickly available and reusable in different contexts. Using such tools, domain experts can directly get the information they need out of their data. However, these tools typically require them to select the visualization type and to specify the visual mappings, which can be difficult because they often lack the necessary visualization knowledge [2]. Knowledge-assisted visualization can address this problem by representing and leveraging formalized visualization knowledge to support the user [3]. Suggesting automatically generated visualizations to the user is one promising approach to aid domain experts in constructing visualizations [2], [4].

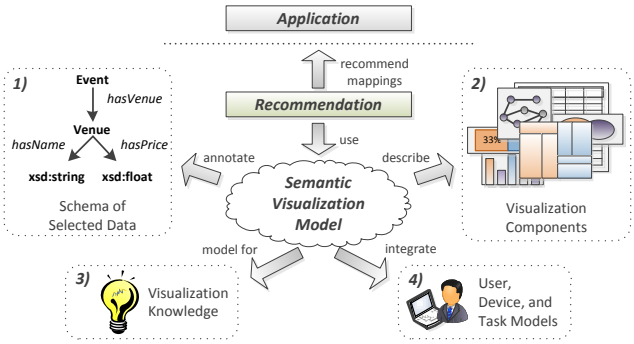


Figure 1. Overview of our goal to recommend mappings of a data source (1) to a visualization component (2) based on a semantic model utilizing visualization knowledge (3) and context information (4).

Consider for example a semantic web data set comprising a list of events hosted at different venues with varying fees. A business user with less visualization experience wants to get an overview of how expensive the events are using his laptop. Thus, he selects a subgraph from a semantic data set as shown in Fig. 1-1 containing two classes (EVENT, VENUE) linked by a Property (*hasVenue*) as well as two Data Properties (*hasName*, *hasPrice*). To map this data to a compatible visualization component (Fig. 1-2), a user needs visualization knowledge (Fig. 1-3). Context information (Fig. 1-4) about the user (knowledge, skills), his device (hard- /software capabilities) and his task (get overview) must also be considered to create a successful mapping. We strive for a generic recommendation approach utilizing and understanding these different ingredients based on a common semantic knowledge model to facilitate the automated visualization process for different tools.

Our goal of creating a knowledge-assisted, context-aware system which recommends visualization components involves a number of challenges, which are addressed by this paper. First, a formalized vocabulary for the interdisciplinary visualization domain is needed. To this end, we have developed a modular visualization ontology called VISO. Second, means to semantically describe visualization characteristics of both data sources and visualization components must be

 ISBN: 978-1-61208-181-6

provided. Therefore, we propose the linking and annotation of semantic web data and component descriptors with concepts of VISO. Third, appropriate visualization components must be discovered for a certain set of data, which includes the deduction of applicable mappings between data and graphic representations. We present a discovery algorithm which takes the aforementioned formalized visualization knowledge and given user requirements into account to search for compatible visualization components. Finally, component candidates need to be ranked with regard to the user, usage and device context, e. g., to consider user language, screen real estate, and available plugins on the device. We have developed a corresponding ranking algorithm for the mappings, i. e., component candidates resulting from the discovery. It explicitly takes into account the visualization knowledge, assigned domain concepts, the user and device context as well as optional criteria from the end user to achieve a context- and task-aware rating. We show the feasibility and practicability of our approach with two prototype implementations: an Eclipse-based visualization tool for semantic multimedia data and a mashup-based visualization workbench called *VizBoard*.

The remainder of this paper is structured as follows. First, we discuss related work in the fields of automated visualization, semantic models for visualization, and semantic-based component recommendation in Section II. Then, Section III introduces our visualization ontology VISO in detail and clarifies how it is applied to describe visualization components and data sources. Afterwards, we present the corresponding recommendation algorithm separated into discovery and ranking in Section IV. Section V gives a brief overview of the prototype implementations and discusses our findings. Finally, in Section VI we conclude the paper and outline future work.

II. Related Work

The recommendation algorithm presented in this paper builds on previous research in the three different research areas (1) automated visualization, (2) semantic visualization models, (3) mechanisms for semantics-based component discovery and ranking. We will now discuss the state of the art in those three areas.

A. Automated Visualization

Several automatic visualization systems have been developed to help users to create visualizations. They produce visualization specifications based on user-selected data and implicitly or explicitly represented visualization knowledge. We distinguish between data-driven, task-driven, and interaction-driven approaches.

Data-driven approaches analyze the meta-model of the data and potentially instance data to generate visualization specifications. Mackinlay addressed the problem of how to automatically generate static 2D visualizations of relational information in his APT system [5]. It searches the design space of all possible visualizations using expressiveness criteria and then ranks them using effectiveness criteria. Gilson et al. developed an algorithm that maps data represented in a domain ontology to visual representation ontologies [6]. Their visual representation ontologies describe single visualization components, e. g., tree maps. A semantic bridging ontology is used to specify the appropriateness of the different mappings. Our automated visualization approach is similar to the one by Gilson et al. in that both data and visualization components are described using ontologies. The main limitation of data-driven approaches is that they do not take other information such as the user's task, preferences or device into account. Task-driven and interaction-driven approaches usually build on the data analysis ideas present in data-driven approaches, but go beyond them.

The effectiveness of a visualization depends on how well it supports the user's task by making it easy to perceive important information. This is addressed by **task-driven approaches**. Casner's BOZ system analyzes task descriptions to generate corresponding visualizations [7]. However, BOZ requires detailed task descriptions formulated in a structured language and is limited to relational data. The SAGE system by Roth and Mattis extends APT to consider the user's goals [8]. It first selects visual techniques based on their expressiveness, then ranks them according to their effectiveness, refines them by adding additional layout constraints (e.g., sorting), and finally integrates multiple visualization techniques if necessary. In constrast to SAGE and BOZ, our algorithm is ontology-based to allow for reasoning and it leverages device and user preference information.

Visual data analysis is an iterative and interactive process in which many visualizations are created, modified and analyzed [2]. **Interaction-driven approaches** consider either the user interaction history or the current visualization state to generate visualizations that support this process. Mackinlay et al. have developed heuristics that use the current visualization state and the data attribute selection to update the current visualization or to show alternative visualizations [9]. Behavior-driven visualization recommendation monitors users' interactions with visualizations, detects patterns in the interaction sequences, and infers visual tasks based on repeated patterns [10]. The current visualization state and the inferred visual task are then used to recommend more suitable visualizations. Interaction-driven approaches leverage implicit state information such as the interaction history, but they consider neither task information that is explicitly expressed by the user, nor user preferences or device constraints.

In summary, while our work builds on many ideas from automated visualization approaches, in particular the work by Gilson et al. [6], it is extensible in terms of visualization components, and it considers task, user preferences and device capabilities. In contrast to generative approaches [5],

 ISBN: 978-1-61208-181-6

[7]–[9], the strength of using visualization components is that such components are optimized for the visual metaphor they represent.

B. Formalizations of Visualization Knowledge

As shown in the previous section, automated visualization requires one or more models to bridge the gap between data and suitable graphic representations. In this regard, prevalent approaches use different concepts, such as rules [8], heuristics [9], and semantic models [6]. We share the view of Gilson et al. [6] that semantic technologies are the methods of choice today. They allow for capturing and formalizing expert knowledge in a readable and understandable manner for humans as well as machines. Therefore, they provide an effective solution for automated recommendation. Further, the current technologies facilitate an easy and dynamic reuse of existing semantic models in new scenarios.

Actually, only few academic works have explored semantic web technologies as means to capture visualization knowledge for describing and recommending resources. Duke et al. [11] were the first proposing the need for a visualization ontology. Their promising approach captures an initial set of concepts and relations of the domain comprising data, visualization techniques, and tasks. Potter and Wright [12] combine formal taxonomies for hard-/software capabilities, sensory experience as well as human actions to characterize a visualization resource. Similarly, Shu et al. [13] use a visualization ontology to annotate and query for visualization web services, with regard to their (1) underlying data model and (2) visualization technique. While the former is a taxonomy comprising various kinds of multidimensional data sets, the latter builds on the data module to classify the graphic representations. For our work, their data taxonomy is not flexible enough as we need to support graph-based data structures for example. Gilson et al. [6] employ three dedicated ontologies to allow for automatic visualization: The first one captures domain semantics and instance data to visualize; the second one describes a particular graphic representation; the final ontology contains expert knowledge to foster the mapping from domain to visualization concepts. In contrast, we allow for a more flexible and generic linking of both sides by annotating each with VISO concepts instead of the explicit, manual creation of an additional ontology. Rhodes et al. [14] aimed to categorize, store and query information about software visualization systems using a visualization ontology as the underlying model. Their approach facilitates methods for specifying data, graphic representation, or the skill of users.

In summary, we share the goal of the works presented above: defining a formalized vocabulary to describe and recommend visualization resources. However, as we strive for a context-aware recommendation we need a more comprehensive and detailed model that covers not only data and graphical aspects, but also represent the user, his activity, and device.

C. Semantics-Based Component Discovery and Ranking

When it comes to finding and binding adequate services for a desired goal, such as visualizing semantic data as we are, *Semantic Web Services* (SWS) tackle a very similar problem. SWS research provides solutions for finding a service or service composition that fulfills a goal or user task based on certain instance data. Therefore, they employ a formal representation of the services' functional and non-function semantics – usually based on description logics – to facilitate reasoning. Based on this, they strive for the automation of the service life-cycle including the discovery, ranking, composition, and execution of services through proper composition environments.

The discovery of suitable semantic services employs either complete semantic service models, e. g., in OWL-S [15] and WSMO [16], or semantic extensions to existing description formats, as proposed by SAWSDL [17] and WSMO-Lite [18]. The former *top-down* approaches are usually very expressive, but descriptions are complex and time-consuming to build. The latter *bottom-up* approaches add semantic annotations, i. e., references to concepts in external ontologies, to WSDL. Even though the above-mentioned solutions cannot be directly applied to our problems, e. g., due to their limitation to web services formats and design principles (stateless), we follow the idea by extending a mashup component description language with semantic references. Thereby, visualization components can be described regarding their data, functional and non-functional semantics, including references to formalized visualization knowledge.

In SWS discovery, suitable services are searched based on a formalized goal or task definition, which is usually a template of an SWS description. Thus, the desired data and functional interface is matched with actual service models. The corresponding algorithms either use measures like text and graph similarities, which restricts the applicability to design-time, or determine the matching degree of services, operations, etc., using logic relationships between annotated concepts as in [19]. In contrast to SWS, we follow a data-driven approach, in which semantically annotated data forms the input for the discovery of suitable candidates. The direct generation of SWS goals from a selected data set is not feasible. Therefore, we individually match data types, functional interface and hard-/software requirements with and between data and visualization components based on shared conceptualizations. Based on this measure, compatible visualization components can be found.

Ranking of service candidates in SWS bears a number of similarities with ranking visualization components for a certain data set. It is usually based on non-functional properties, such as QoS and context information (user profile, device

 ISBN: 978-1-61208-181-6

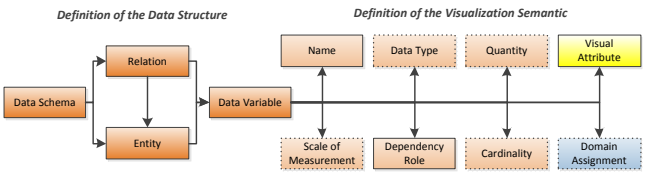


Figure 2. Overview of the VISO data module

capabilities). To this end, a number of sophisticated concepts exist, e. g., for multi-criteria ranking based on semantic descriptions of non-functional service properties [16] and for context sensitive ranking [20]. Since these algorithms are rather generic and work on a semantic, non-functional level, they likewise apply to our concept space.

In summary, the discovery and ranking of candidate services for a predefined goal in SWS research follows a similar principle as our work. Yet, its solutions can not be directly applied to our problems. For one, there is a difference in component models, e. g., with regard to statefulness of visualization components. Furthermore, the discovery of visualization components can not be based on predefined, formalized goal descriptions, as it basically depends on semantic data which is annotated with visualization knowledge. For the annotation of visualization components with semantic concepts though, we can apply the ideas of SAWSDL and WSMO-Lite to the component descriptions. To *link* semantic data with visualization components, a shared conceptualization of visualization knowledge is needed. Therefore, the next section presents VISO.

III. VISO: A Modular Visualization Ontology

The foundation of our visualization recommendation approach is a formalized, modular visualization ontology called VISO [21]. It provides a RDF-S/OWL vocabulary for annotating data sources and visualization components, contains factual knowledge of the visualization domain, and serves as a semantic framework for storing contextual information. Altogether, it serves as a *bridging ontology* between semantic data and visualization components by offering shared conceptualizations for all four mapping ingredients shown in Fig. 1. Details of VISO and its development are described in [21]. The seven VISO modules (data, graphic, activity, user, system, domain, and facts) represent different facets of data visualization domain. They refer to each other and to existing ontologies as needed. VISO modules can be extended to accommodate new concepts.

1) Data: Fig. 2 shows the data module which contains concepts for describing data variables and structures for visualization purposes. While all concepts are employed to describe visualization components, those with dotted lines are also used to annotated semantic data. The vocabulary is especially need at component-side to describe possible input data in a generic manner as the most of visualizations allow for representing domain independent data. For example, a simple table may visualize data about hotels, cars, or humans. Using this vocabulary, we specify only the data structure and characteristics. As can be seen, a DATA SCHEMA consists of ENTITY and RELATION concepts. The latter represent links between ENTITY concepts like an OWL Object Property. Both ENTITIES and RELATIONS can contain DATA VARIABLE concepts, whose equivalent in OWL space is a Data Property. For example, the semantic data model of a table visualization component would be represented as one ENTITY concept with several DATA VARIABLES for every column. Further semantics, e. g., the SCALE OF MEASUREMENT and CARDINALITIES – specified using built-in OWL constraints – can be defined on the DATA VARIABLE concepts (cf. Fig. 2) to constrain its, e. g., its scale. By linking the concepts from the data module to the VISUAL ATTRIBUTE concepts from the graphic module, we bridge the gap between data attribute and visual elements and properties.

2) Graphics: The graphics module conceptualizes the semantics of GRAPHICAL REPRESENTATIONS and their parts, e. g., their VISUAL ATTRIBUTES. Concrete graphical representations, e. g., *scatter plot* and *treemaps*, and concrete visual attributes such as *hue* or *shape* are contained as instance data. The concepts from the graphics module are used to semantically annotate visualization components and to define visualization knowledge in the facts module.

3) Activity: The activity module models user activity in a visualization context. It builds on the ontology-based task model by Tietz et al. [22], which distinguishes betweens high-level, domain specific TASKS and low-level, generic ACTIONS, similar to the distinction made by Gotz and Zhou [23]. We have extended the action taxonomy of Tietz's task model by separating data- and UI-driven ACTIONS, and by formalizing ACTIONS from the visualization literature such as *zoom* and *filter*. This enables the fine-grained annotation of interaction functionality in visualization components.

4) User: The user module formalizes user PREFERENCES and KNOWLEDGE. Users can, for example, have PREFERENCES for different GRAPHICAL REPRESENTATIONS, and their *visual literacy* can differ. As manifold context models for users, their characteristics and preferences, already exist those can be seamlessly integrated and used here.

5) System: The system module facilitates the description of the device context, e. g., installed PLUG-INS or SCREEN SIZE. It also allows us to annotate a visualization component with its system requirements. Again, sophisticated models for device characteristics and context exist, which were reused or integrated in this module. As an example, we borrow concepts from the *CroCo* ontology [24], which combines user, usage, system, and situational context from different existing works developed by academia.

6) Domain: Many visualizations are domain-specific, and thus it is important to consider the domain context during

 ISBN: 978-1-61208-181-6

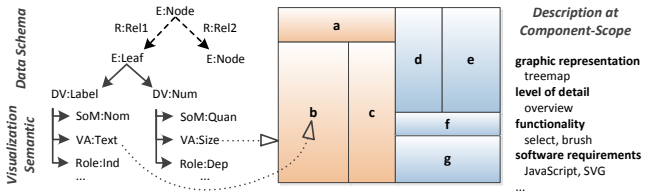


Figure 3. Description of a treemap visualization in VISO.

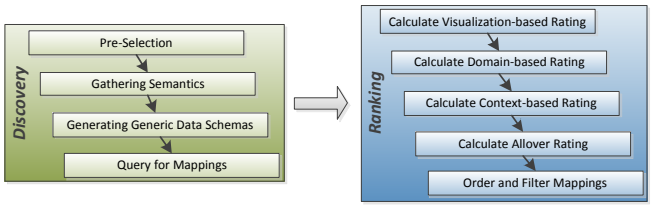


Figure 4. Overview of the recommendation algorithm.

visualization recommendation. However, it is not feasible to model all possible visualization domains. Instead, we support linking to existing domain ontologies. A DOMAIN ASSIGNMENT links VISO concepts, e. g., a DATA VARIABLE (cf. Fig. 2), to concepts from specific domain ontologies. As this assignment is usually created automatically during data analysis, it can be qualified with a probability value reflecting its accuracy. Thus, the analysis of a data source with ambiguous Properties, such as *typeOfJaguar* and *typeofApple*, will result in multiple domain assignments with probabilities below 1. In contrast, a Data Property *hasPrice* from our motivating example could be annotated with *price* and a probability of 1. A visualization component supporting DATA VARIABLE annotated with the more general concept *value* could be inferred as a possible mapping.

7) Facts: The visualization recommendation also depends on factual visualization knowledge to select suitable visualizations. Thus, we formalized knowledge from the information visualization community, e. g., verified statements such as "position is more accurate to visualize quantitative data than color" [25], to make it machine-processable. These rankings and constraints are formalized in rules in the Facts module. These rules use of the vocabulary of the other VISO modules in their conditions part, e. g., SCALE OF MEASUREMENT (*quantitative*) and the VISUAL ATTRIBUTE (*position, color*) for the mentioned example. If the conditions are matched, a rating is assigned to the corresponding visualization component description.

To give a more practical insight, the following example explains how a *treemap* visualization is described using VISO (see Fig. 3). First, the hierarchical data structure of the *treemap* is specified. At the top level, a *Node* ENTITY represents the whole *treemap*. It can contain *Leaf* ENTITIES and *Node* ENTITIES. The label and size variables of *Leafs* can be configured. They are annotated with visualization semantics, e. g., the SCALE OF MEASUREMENT for the label variable is *nominal* and the ROLE of the size variable is *dependent*. Further domain semantics could be added to the variables, e. g., WordNet (http://wordnet.princeton.edu/) concepts such as *value*. In addition to the data structure and the variables, more general semantics such as the kind of GRAPHIC REPRESENTATION (*treemap*), the LEVEL OF DETAIL (*overview*) and possible ACTIONS (*select*, *brush*) are defined for the entire visualization component.

In order to facilitate the construction of visual mappings, VISO is used to annotate visualization components and semantic web data. In the latter case, we annotate only RDF Properties on a schema level. RDF Properties hold the data that will be visualized, e. g., literals and relations, whereas RDF classes assemble such properties and do not provide additional information that would be relevant for visualization. Similarly, annotations are made on the schema level, because instance data annotation would be redundant. Consider our motivating example (see Fig. 1-1), comprising the Property *hasPrice*. Because the Property has the RDFS Range *xsd:float*, the required DATA TYPE is already defined and the SCALE OF MEASUREMENT is *quantitative*. The number of distinct values (CARDINALITY) and the overall number of values (QUANTITY) can be extracted from the instance data. While a DOMAIN ASSIGNMENT is not mandatory, it could be applied, e. g., to *price* from the WordNet vocabulary.

In summary, VISO models the concepts required for data visualization. It is used to annotate data, to describe visualization components, to represent context and factual knowledge. Together, these different pieces are the foundation of our visualization recommendation algorithm.

IV. VISUALIZATION RECOMMENDATION ALGORITHM

The visualization recommendation algorithm creates an ordered list of mappings to visualizations components for the selected data (see Fig. 1-1). It considers contextual information (e.g., device, user model) as well as knowledge about the full data source. While the user model and device are mandatory inputs, visualization specific information like the required LEVEL OF DETAIL or the requested kind of GRAPHICAL REPRESENTATION can be provided as optional constraints.

The algorithm consists of two separate steps: discovery and ranking (see Fig. 4). Both steps leverage semantic knowledge formulated as VISO concepts (see Section III). In the discovery step, potential mappings between data and widgets are generated based on functional requirements. The resulting visualization set is then sorted in the ranking step using the formalized visualization knowledge and domain concepts, as well as by contextual and visualization specific information.

 ISBN: 978-1-61208-181-6

A. Discovery of Mappings

The discovery algorithm generates a set of mappings from the selected data to visualization components (see Fig. 4). First, potentially applicable widgets are identified and non-applicable components are ruled out (**pre-selection**), since limiting the set of available visualization components early improves the overall algorithm performance. To be applicable, a widget has to (1) be compatible with the target device (e. g., required PLUGINS must be available), (2) support the number of selected Data Properties, and (3) support visualization and task specific requirements (e. g., showing an *overview*), if specified by the user. As can be seen, these constraints don't relate to data structure or semantics of the data variables, yet. Semantic matching is carried out with the resulting component candidates in the following step.

Second, semantics, e. g., the SCALE OF MEASUREMENT, DATA TYPE, and QUANTITY (Fig. 2) of the selected Properties are fetched (**gathering semantics**). For example, the DATA TYPE *xsd:float* or the SCALE OF MEASUREMENT *quantitative* of the property *hasPrice* (see Fig. 5-3)) would get retrieved. This semantic information about the Properties is used in the next steps.

Third, we **generate generic data schemas**, which are then used to query for mappings. We distinguish between tabular and graph-based DATA SCHEMAS. TABULAR DATA SCHEMAS contain one ENTITY with several DATA VARIABLES (Fig. 5-1). GRAPH-BASED DATA SCHEMAS contain two or more linked ENTITIES, each containing zero or more variables (Fig. 5-2).

If a **single class** has been selected, a TABULAR DATA SCHEMA is chosen and an ENTITY is created for that class. For every selected Data Property of this class, a DATA VARIABLE with the semantic information (that was retrieved in the previous step) is attached to the ENTITY.

If **several classes** have been selected, we generate both a tabular and a graph-based DATA SCHEMA. For the TABULAR DATA SCHEMA, a single ENTITY gets created. For any selected Data Property from those classes, a DATA VARIABLE with the semantic information is attached to the single ENTITY. This reduces the graph-based data structure to a tabular structure. For example, consider the data shown in Fig. 5-3. The algorithm would create one ENTITY with two DATA VARIABLES. The first DATA VARIABLE would represent the semantics of *hasName*, e. g., the *nominal* SCALE OF MEASUREMENT, and the second DATA VARIABLE would represent *hasPrice*. The GRAPH-BASED DATA SCHEMA gets generated as follows: Beginning with a class from the input data, e. g., *Event* in Fig. 5-3, an ENTITY is created. Similar to the other cases, DATA VARIABLES and their semantics are attached to this ENTITY for the selected Data Properties linked to the class. Next, for each Object Property connected with the class, a RELATION gets generated. If the target class for that RELATION has not been processed yet, it is created and processed in a similar way. This depth-first processing continues until the current part of the input graph is completely traversed. If there are multiple unconnected classes in the input, the algorithm continues with those until all graph components are processed. For example, the algorithm would generate the DATA SCHEMA illustrated in Fig. 5-4 by processing the input data structure shown in Fig. 5-3.

Fourth, the mappings are generated by querying the semantic representations of the pre-selected components with the generic DATA SCHEMAS that were computed in the previous step (**query for mappings**). The mappings include permutations of DATA VARIABLES with similar semantics, and thus the number of mappings may be higher than the number of existing components. Using the data structure generated by the algorithm for the example shown in Fig. 5-4, both the *scatter plot* (Fig. 5-1) and the *treemap* (Fig. 5-2) would fit on the level of data structure. However, only the *treemap* is a suitable mapping due to the annotated semantics which are also employed by querying. The *scatter plot* is not suitable because it has two *quantitative* DATA VARIABLES where both a *nominal* and a *quantitative* DATA VARIABLE are required. The generated set of mappings from the selected data to the visualization components is ranked in the next part of the algorithm.

B. Ranking of Mappings

The ranking step of the algorithm sorts the visual mappings that were generated by the previous discovery step. While the discovery step identifies valid mappings and visualization components that satisfy functional criteria, it does not take their effectiveness into account. To sort the mappings by their effectiveness, the ranking step applies factual visualization knowledge, domain assignments and contextual user and device information.

1) Factual Visualization Knowledge: The factual visualization knowledge (see Section III) is defined by a set of rules which consist of a condition and a rating. The conditions are specified using the VISO vocabulary for the visualization components. For each widget, the ratings of all rules that are met are added to its specification. During runtime, the arithmetic mean of all ratings r_{v_i} is calculated for the discovered component of each visual mapping. For example, we formalized rules to rate the appropriateness of visual encodings for quantitative data [25]. The *quantitative* DATA VARIABLE of the *treemap* (Fig. 5-2) is rated with *0.5* as it employs "only" *size* and not *position*.

2) Domain Assignments: Domain concepts from various ontologies are assigned to both the data input and the visualization components with a certainty value (see Section III). For each pair of input Property and DATA VARIABLE of the visualization component, we calculate a semantic similarity rating between 0 and 1 (e. g., using [26]), if they both have a domain concept assigned with a certainty greater

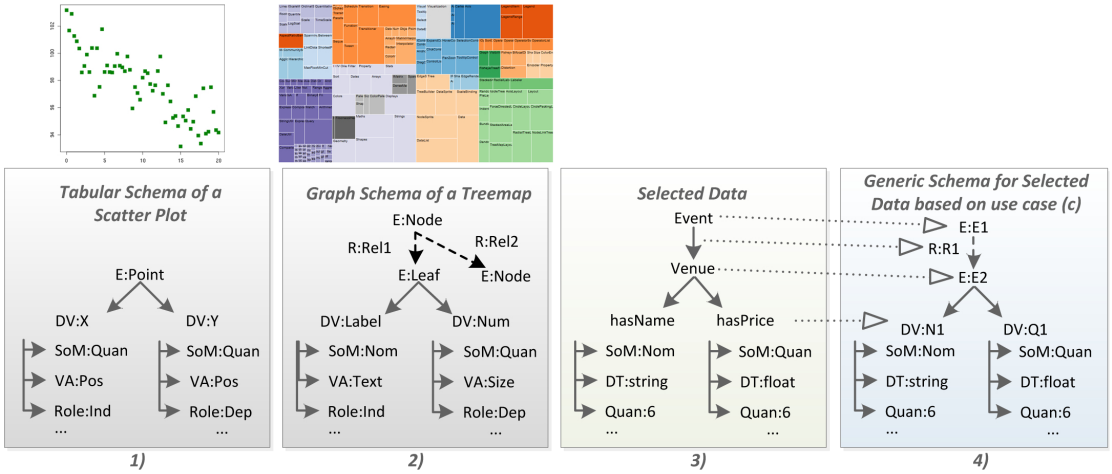


Figure 5. Comparison of the data structure and the annotation between 1) a scatter plot, 2) a treemap, 3) user's selected data, and 4) a generic equivalent of the selected data.

than 0. The final rating r_{d_j} is the product of the semantic similarity and the arithmetic mean of both certainties. In our example (see Sect. III), we used *value* and *price* from WordNet to annotate the *quantitative* DATA VARIABLE of the *treemap* and the Property *hasPrice* from our data set, each with a certainty of *1*. Using [26], we get a rating r_d=0.9094.

3) User and Device Information: The rules for the context-based rating r_c are part of the knowledge base and use the VISO vocabulary, similar to the factual visualization knowledge. The rules are executed during runtime and employ the above mentioned identifiers of users' and their device context models. For example, we construct a SPARQL-based rule that counts the use of different GRAPHIC REPRESENTATIONS, like *treemaps* or *scatter plots*. This rule assigns a rating r_{c_k} between 0 and 1 to the visual mappings.

The three different kinds of rating are combined using an arithmetic mean. The overall rating has a range between 0 and 1. We weight all three rating types equivalently for two reasons. First, the assignment of a (quantitative) rating is often subjective. Second, a profound user study is needed to evaluate the impact of each knowledge base in users visualization selection process what will be future work. As x, y, and z are the number of each kinds of rating, the overall rating R for each mapping is calculated in terms of

$$R = \frac{1}{3}\left(\frac{1}{x}\sum_{i=1}^{x} r_{v_i} + \frac{1}{y}\sum_{j=1}^{y} r_{d_j} + \frac{1}{z}\sum_{k=1}^{z} r_{c_k}\right)$$

The list is ordered based on the combined ratings R for each mappings. This ranking could be used to automatically display the top mapping to the user as a visualization, or, as in our approach, to let the user pick one of the top n ranked visualizations. We next discuss our implementation of the visualization recommendation algorithm in two research projects.

V. IMPLEMENTATION AND DISCUSSION

To realize the concepts discussed above, we first developed VISO as an open, modular ontology (available at http://purl.org/viso/) based on OWL DL. It is comprised of concepts, properties and instance data from the visualization domain, as well as factual knowledge modeled using Jena Rules (http://jena.sf.net/). Details on the design process and decisions can be found at [21].

We then implemented our generic recommendation approach and integrated it with two existing research projects: *KIMM* [27] and *CRUISe* [28]. The algorithms build on Jena to manage all semantic models and employ SPARQL 1.1 to query the knowledge bases and to create the mappings.

Within the frame of *K-IMM*, we enhanced the Eclipse RCP-based application *Sim²* which allows for visualizing RDF-based multimedia data. With a wizard we enable users to select classes and properties from theirdata set for visualization. Since all views are semantically annotated with concepts from VISO, this input can be used for our discovery algorithm, which recommends suitable visualizations for the selected data. Since *Sim²* neither tracks nor uses any context information, this integration was limited to the discovery. The context-aware rating was omitted in this prototype. In this prototype, the discovery mechanism, which relies only on functional and objective matching, was able to identify suitable visualizations.

Our approach is also an integral part of *VizBoard*, an information visualization workbench for semantic data based on the mashup platform *CRUISe*. *CRUISe* facilitates the dynamic, context-aware composition of mashups from distributed web resources. Hence, it builds on a universal component model which includes semantic descriptors. We

 ISBN: 978-1-61208-181-6

encapsulated a number of well-known visualizations, including several from the Protovis library [29], and employed the descriptors to realize the annotation with VISO concepts. The discovery and ranking algorithms were integrated as part of a multi-step wizard, which results in the context-aware recommendation of suitable visualization components as basis for a mashup UI. Using this implementation, we could also evaluate our ranking algorithm. Even though it performs as expected, more work is needed for two reasons. First, we build on visualization knowledge, e. g., the expressiveness of visual attributes for quantitative data, that reflects the current state of knowledge in the field of information visualization. However, due to limited empirical evidence and different expert opinion, some of the current guidelines are not accepted and could change in the future. Second, user studies to identify the impact created by the visualization, domain, and context knowledge in the visualization process could improve the weighting of these three components.

As with many search mechanisms, defining the goal of the visualization selection process is challenging. We argue that the query creation should allow for defining requirements and options, e. g., "I like to visualize A and B, and C if possible", in an uncomplicated way. Hence, a sophisticated user interface should assist the user during the goal definition. Furthermore, our discovery algorithm needs to be extended to support optional input.

A limitation of semantic approaches like ours is the need of descriptions respectively of annotations. Thus, it is up to component authors and data providers to augment their components/data using the VISO concepts correctly. In this regard, adequate tool support would be beneficial.

VI. Conclusion and Further Work

Selecting an appropriate visualization for a specific data set in a specific scenario remains challenging for non-experts. Therefore, we have presented a context-aware and knowledge-assisted approach to recommend suitable visualizations for semantic web data. Its foundation is the modular visualization ontology VISO which provides the vocabulary to annotate both data sources and visualization components. Based on these shared concepts from the visualization domain, our recommendation algorithm covers both discovery and context-aware ranking of suitable graphic representations: First, possible mappings from data to visual encodings are identified using the selected data, its semantics, and other functional information. Then, quantitative ratings for each mapping are calculated with respect to visualization knowledge, domain concept relations and context information.

As our implementations shows that the approach can recommend visualization components based on semantics from different sources, the discussion shows some directions for future work. We will investigate a tool for the semi-automatic annotation of visualization semantics for semantic web data. We are also planning to conduct a user study to identify and model the interdependencies between the knowledge bases employed within the ranking. To enhance users interactive selection of data, task- and visualization-specific input for the algorithm, we are working on a faceted browser which will distinguish between requirements and weighted optional criteria.

Acknowledgment

This work is partly funded by the German Federal Ministry of Education and Research under promotional reference number 01IA09001C.

The authors wish to thank Michael Aleythe for implementation support as well as David Rusk and Patrick Gorman for helpful editing suggestions.

References

[1] R. Haber and D. A. McNabb, "Visualization idioms: A conceptual model for scientific visualization systems," *Visualization in Scientific Computing*, pp. 74–93, 1990.

[2] L. Grammel, M. Tory, and M.-A. Storey, "How information visualization novices construct visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, pp. 943–952, November 2010.

[3] M. Chen, D. Ebert, H. Hagen, R. Laramee, R. van Liere, K.-L. Ma, W. Ribarsky, G. Scheuermann, and D. Silver, "Data, information, and knowledge in visualization," *Computer Graphics and Applications, IEEE*, vol. 29, no. 1, pp. 12 –19, Jan. 2009.

[4] J. Heer, F. van Ham, S. Carpendale, C. Weaver, and P. Isenberg, *Creation and Collaboration: Engaging New Audiences for Information Visualization*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 92–133.

[5] J. Mackinlay, "Automating the design of graphical presentations of relational information," *ACM Trans. Graph.*, vol. 5, no. 2, pp. 110–141, 1986.

[6] O. Gilson, N. Silva, P. Grant, and M. Chen, "From web data to visualization via ontology mapping," in *Computer Graphics Forum*, vol. 27, no. 3. Blackwell Publishing Ltd, Sep 2008, pp. 959–966.

[7] S. M. Casner, "Task-analytic approach to the automated design of graphic presentations," *ACM Trans. Graph.*, vol. 10, pp. 111–151, April 1991.

[8] S. F. Roth and J. Mattis, "Automating the presentation of information," in *Artificial Intelligence Applications, 1991. Proc. of 7th IEEE Conf. on*, vol. 1. IEEE, 1991, pp. 90–97.

[9] J. Mackinlay, P. Hanrahan, and C. Stolte, "Show me: Automatic presentation for visual analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1137–1144, Nov. 2007.

[10] D. Gotz and Z. Wen, "Behavior-driven visualization recommendation," in *IUI '09: Proceedings of the 13th international conference on Intelligent user interfaces*. New York, NY, USA: ACM, 2009, pp. 315–324.

 ISBN: 978-1-61208-181-6

[11] D. Duke, K. Brodlie, D. Duce, and I. Herman, "Do you see what i mean? [data visualization]," *Computer Graphics and Applications, IEEE*, vol. 25, no. 3, pp. 6–9, May 2005.

[12] R. Potter and H. Wright, "An ontological approach to visualization resource management," in *DSV-IS*, 2006, pp. 151–156.

[13] G. Shu, N. Avis, and O. Rana, "Bringing semantics to visualization services," *Advances in Engineering Software*, vol. 39, no. 6, pp. 514–520, 2008.

[14] P. Rhodes, E. Kraemer, and B. Reed, "Vision: an interactive visualization ontology," in *ACM-SE 44: Proceedings of the 44th annual Southeast regional conference*. New York, NY, USA: ACM, 2006, pp. 405–410.

[15] D. Martin, M. Burstein, D. McDermott, S. McIlraith, M. Paolucci, K. Sycara, D. McGuinness, E. Sirin, and N. Srinivasan, "Bringing semantics to web services with owl-s," *World Wide Web*, vol. 10, pp. 243–277, Sep. 2007.

[16] D. Fensel, M. Kerrigan, and M. Zaremba, *Implementing Semantic Web Services: The SESA Framework*. Springer, 2008.

[17] J. Farrell and H. Lausen, "Semantic annotations for WSDL and XML Schema," http://www.w3.org/TR/sawsdl/, W3C, Aug. 2007.

[18] J. Kopecký and T. Vitvar, "Wsmo-lite: Lowering the semantic web services barrier with modular and light-weight annotations," in *Proc. of the Intl. Conf. on Semantic Computing*, Aug. 2008, pp. 238–244.

[19] Y. Chabeb, S. Tata, and A. Ozanne, "YASA-M: A semantic web service matchmaker," *Proc. of the Intl. Conf. on Advanced Information Networking and Applications*, pp. 966–973, Apr. 2010.

[20] F. Gilles, V. Hoyer, T. Janner, and K. Stanoevska-Slabeva, "Lightweight composition of ad-hoc enterprise-class applications with context-aware enterprise mashups," in *Proc. of the Intl. Conf. on Service-Oriented Computing*. Springer, 2009, pp. 509–519.

[21] M. Voigt and J. Polowinski, "Towards a unifying visualization ontology," TU Dresden, Institut fuer Software und Multimediatechnik, Dresden, Germany, Technical Report TUD-FI11-01, Mar. 2011, ISSN: 1430-211X.

[22] V. Tietz, G. Blichmann, S. Pietschmann, and K. Meiner, "Task-based recommendation of mashup components," in *Proceedings of the 3rd International Workshop on Lightweight Integration on the Web (ComposableWeb 2011)*. Springer, Jun. 2011.

[23] D. Gotz and M. Zhou, "Characterizing users' visual analytic activity for insight provenance," in *Visual Analytics Science and Technology, 2008. VAST '08. IEEE Symposium on*, Oct. 2008, pp. 123 –130.

[24] A. Mitschick, S. Pietschmann, and K. Meißner, "An ontology-based, cross-application context modeling and management service," *Intl. Journal on Semantic Web and Information Systems (IJSWIS)*, Feb. 2010.

[25] W. S. Cleveland and R. McGill, "Graphical perception: Theory, experimentation, and application to the development of graphical methods," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 531–554, 1984.

[26] D. Lin, "An information-theoretic definition of similarity," in *Proc. 15th International Conf. on Machine Learning*. Morgan Kaufmann, San Francisco, CA, 1998, pp. 296–304.

[27] A. Mitschick and K. Meissner, "Generation and maintenance of semantic metadata for personal multimedia document management," in *Advances in Multimedia, 2009. MMEDIA '09. First International Conference on*, july 2009, pp. 74 – 79.

[28] S. Pietschmann, "A model-driven development process and runtime platform for adaptive composite web applications," *International Journal on Advances in Internet Technology (IntTech)*, vol. 4, no. 1, pp. 277–288, February 2009.

[29] M. Bostock and J. Heer, "Protovis: A graphical toolkit for visualization," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 15, no. 6, pp. 1121 –1128, nov.-dec. 2009.

 ISBN: 978-1-61208-181-6

Reflective Case-Writing Environment using a Multi-representation Schema for Medical Service Education

Wei Chen, Masaki Fujii, Liang Cui, Mitsuru Ikeda
School of Knowledge Science
Japan Advanced Institute of Science and Technology
Nomi, Japan
{wei.chen; masaki.fujii; cui-liang; ikeda}@jaist.ac.jp

Noriyuki Matsuda
Department of Computer and Communication Sciences,
Faculty of Systems Engineering
Wakayama University
Wakayama, Japan
matsuda@sys.wakayama-u.ac.jp

***Abstract* — In this research, by developing a learning support system for medical services, we will establish an approach that supports medical profession novices to improve proficiency understanding patient-centered medical services. Using an ontology in this paper, as a first step of the project, we organized a learning model which promotes reflective learning of the case-method for medical service education. As an implementation of the learning model, we established a learning environment that support learners to reflect on their thinking process in their experiences by a learning strategy which consists of three case-writing phases: the description phase, the cognitive conflict phase, and the knowledge building phase.**

Keywords-Thinking skill; Case-Method; Ontology; Medical Service Education.

I. INTRODUCTION

There are always many difficult problems continually appearing from various viewpoints in medical practice. Actually, medical staff always feels a vague anxiety that their dedicated efforts could not catch up with the increasing problems. Moreover, to provide high quality medical services that can respond to the various and high-degree increasing patients' demands is becoming an important and urgent issue in medical service practice. The subjects in medical service education in a broad sense include both the medical knowledge/skills for the medical diagnosis or the treatment and the interpersonal skill to facilitate the prompt and smooth implementation of medical services. In this research, we focus on the latter as the matter of medical service sciences in a narrow sense, while we address the former as the matter of "medical education" and will not be deeply involved in it.

We believe that the service science approach can make a contribution to establish a methodology to improve the quality of Medical Services in a narrow sense. One of the pioneers in the field of Service Science, Yoshikawa has proposed that the model for service improvement is that the knowledge circulation of intellectual collaboration by the persons concerned in the service promotes to create and refine the service knowledge. Moreover, he implies that the knowledge circulation will cause the ideal of societal innovation [1]. In the medical viewpoint, we think it is necessary to refine the education approaches for supporting the medical knowledge circulation by improving the medical practitioners' thinking ability to collaboratively create and refine a medical service knowledge.

In this research, by developing the learning support system for medical services, we will establish an approach that supports the medical profession novices to improve their proficiency in understanding patient-centered medical services. The current goal of this research is to make a rational learning model for medical service education and try to establish a methodology to create design loop for medical service educational program development but not to make strong contributions to technological medical service education.

II. DIFFICULTIES IN MEDICAL SERVICE EDUCATION

In recent medical practice, the traditional apprenticeship-style on-the-job training system, so-called of, "seniors train novices strictly on the job" is vanishing gradually because of mental resistance for novices to accept the evidence-lacking, experience-based guidance of implicit medical service knowledge from seniors. Moreover, newcomers who have poor insight and sensitivity to people are increasing, and there appears to be an increasingly pronounced tendency for the medical staff to be unable to learn medical service knowledge or skills to understand patients' minds through communication with other medical staff.

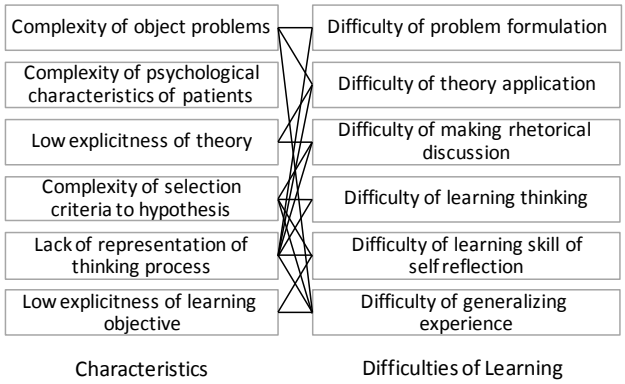


Figure 1. Characteristics of medical service and difficulties of learning the knowledge required for it

For example, when a novice nurse takes charge of pediatrics, he may be puzzled by the complexity of emotional engagement between the child patients who are weaker than himself, their parents who are exceptionally

 ISBN: 978-1-61208-181-6

anxious about their children's health and the doctors who conduct medical treatment. In order to have an acute insight into the complex structure of emotional engagement, it is necessary to have a rich sensitivity for understanding others' minds and, a rational attitude of the acceptance of and respect for the immature hearts of the pediatric patients. That is typical tacit knowledge which is not easy to acquire for novice medical staff.

For the purpose of developing medical human resources with higher cognitive ability as shown in Figure 1, a variety of educational methods to foster the tacit knowledge or tacit skill by coaching the thinking process has been offered to the medical staff. For example, in the field of nursing education, teaching approaches such as clinical conferences, reflective journals, narrative methods, case-method, etc. are conducted on a routine basis at many hospitals. However, in such a practical learning environment, it is said that the major difference between the learners who can learn what should be learned and the learners who cannot learn very well comes from differences in learners' sensibility or insight to others' minds. Moreover, even though learners have successfully taught tacit knowledge in the practical learning environment, most of them face more serious difficulties to assimilate the knowledge to their own existing knowledge and organize it as general knowledge to be applicable to future similar situations. The difficulties are caused by lack of the experience of making "thinking about others' minds" a subject for meta-level logical thinking, while most people guess others' minds only by intuition. Therefore, to foster the ability of meta-level logical thinking seems to be accompanied by an essential difficulty caused by the essential nature of humanity. In addition, the complexity of the matters of minds, the low explicitness of theory, the complexity of selection criteria for hypotheses, a lack of representation of thinking process, etc., make it difficult for novices to learn the knowledge required for medical services (Figure 2).

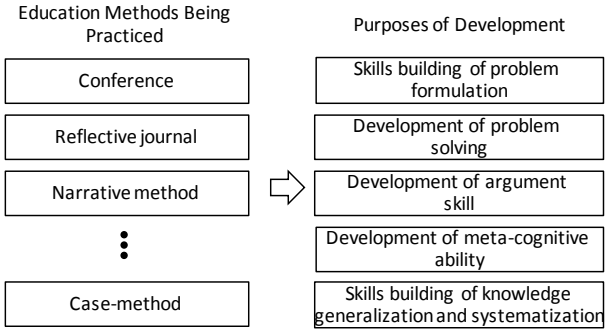


Figure 2. Fostering tacit knowledge/skills by coaching-thinking process

In this research, we focus on the case-method as an approach to Medical Service Education. One of the educational principles behind the case-method in business management education is "if you want to teach how to deal with a new problem that you have not yet experienced, we should teach them how to think. In fact, the ability of thinking about thinking and the ability of dealing with new problems can be regarded as different issues in principle but they are completely the same issue in practice." [2].

III. OVERVIEW OF CASE-METHOD

In the case-method, in order to acquire "skills to deal with new problems that have not been experienced yet", the learners are assigned the task to think how to deal with the "real" problems that have occurred in their own practices and write their thoughts and behavior to cope with the problems as "cases". Then, they join a group discussion on the case with other learners to investigate the validity of their own thinking process from various viewpoints and co-create new solutions to the "real" problem. Through these learning experiences, they learn the learning ability to deal with highly-non-deterministic and highly-complex practical problems [2].

The actual flows of the case-method in business management education are as follows: (1) the instructor distributes the prepared case materials to the learners in advance. (2) The learners organize the contents of the case to analyze and identify the core issues. The analysis should be made based on the facts in the case, the assertion inferred from the facts, insight into the thinking processes of the agents in the case, and the learners' own knowledge. (3) According to the analysis, they think out their own solutions to the problem. After that, (4) the learners join the discussion on the validity of each learner's solution, where the instructor will not join the discussion actively but just raise the topic to be discussed and lead the flow of the discussion [3].

When designing learning materials for the case-method, it is necessary to (1) write down the events that actually occurred, (2) to consider how the learners think about the case and how they will discuss it. Therefore, it is essential for a case-writer to be able to estimate how learners think or how their discussion goes on from the deep understanding of written issues on the case [4].

A. Learning in Case-Method

In the survey paper on the argument study, Maruno and Tomita [5] claim that most researchers focus on the argumentative skills to examine the rationality or validity of information or knowledge used in the discussion. On the other hand, the skills to produce or externalize ideas in the discussion have not been studied in the research field. However, based on the empirical and the theoretical research so far, the former skills cannot be acquired without the latter ability. It implies that by participating in activities in which the latter skills in required repeatedly, the former skills can be acquired.

Moreover, they support the Kuhn(1991) model of internal thinking process as a dynamic internal dialogue base on Billig's idea that "people engaged in problem solving or decision making, try to make the best judgment of selecting one from some possible options by justifying each of them from many different viewpoints and comparing the justifications to the options" [6] [7]. The reason why they strongly rely on Kuhn's model is that the model shows clear socio-cultural explanation of how the argument guides the

 ISBN: 978-1-61208-181-6

thinking process, which is, one regards the thinking developing process as a more dynamic and clarifies the tight relationship between individual internal process of thinking and social process of thinking such as exchanging position with others and the individual process.

Standing on this viewpoint, the case-method can be used as a concrete educational approach for learning internal dialogue. On the other hand, it is difficult to learn the dynamic internal dialogue associated with social interaction for the reason (shown in Figure 1) that particularly higher cognitive ability is required. In our research project, in parallel, we have been developing an educational program that can reduce the learner's load in learning the association between internal dialogue [8] and social interaction [9].

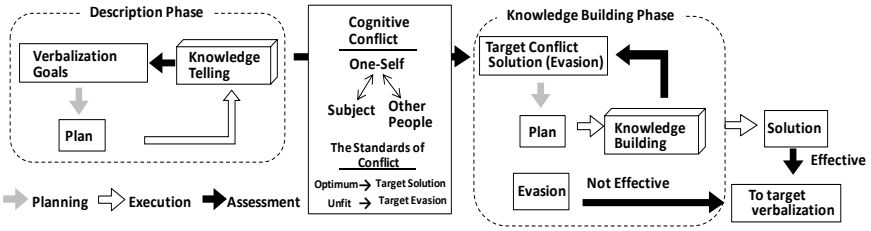


Figure 3. A goal-attainment model of verbalization as a learning strategy

B. *Learning by Designing Case Learning Materials*

Ito proposed, by analyzing of effect of the verbalization as a learning strategy, a model of learning goals achievement by verbalization as an integrated model of three learning mechanisms, that is, tutoring that focuses on the learning effect of the teaching activities, self-explanatory quality (nature?) of learning activities, and collaborative learning among learners [10].

We believe that learners can be active entities who can find a meaningful entity for the goal of knowledge acquisition by themselves, and they can achieve the goal by externalizing their self-explanatory of their thinking process to other learners. The externalization processes consist of the two phases of the knowledge description phase, and knowledge building phase and the cognitive conflict can be bridging activities of the two phases as shown in Figure 3. We will discuss the three phases in detail below.

The description phase is an iteration of the internal learning activities to achieve the goal of verbalization by externalizing one's thought in his own experiences. The cognitive conflict is a trigger cognitive process for learners to go into the knowledge building phase by facing the conflict states (realization of cognitive gap among learners' mental models, cognitive differences with other learners, or errors in their knowledge) through the verbalization of their thought and interaction with others. And then, in the knowledge building phase, the learners aim at achieving the goal of resolving those conflict states. The goal of verbalization in the knowledge building phase is to resolve the conflicts and is essentially different from the goal of verbalization in the knowledge description phase. This goal achievement model can be regarded as a learning model that includes the model of thought for dynamic internal dialogue mentioned above.

As mentioned at the beginning of this chapter, the design of case materials requires: (1) writing the case, (2) preparing the content that should be considered and the set of branch points for discussions. In this research, we aim at developing learners' meta-cognitive skills by imposing the design tasks of case-method learning materials on the learners and promoting cognitive interaction with others.

In particular, as an educational program for the medical professions (the nurses in this paper), we developed a learning environment for realizing a model of learning goals achievement by verbalization. Using the environment, the nurses write down (the description) their own thinking process in their experience as cases, guess others' different thoughts, find a cognitive conflict from the thoughts and try to resolve the conflicts by building new knowledge [11] [12].

IV. Environment Supporting Learning in Design Learning Materials

In Figure 3, in the learning strategy, learners engage with verbalization activities in the description phase and the knowledge building phase, and the activities are externally observable at the behavioral level. Meanwhile, the activities of making goals, plans, cognitive conflicts, resolving conflicts etc., are not externally observable internal cognitive activities.

Since those activities are relatively abstract and ambiguous, it is difficult for the learners to achieve the learning goals. The difficulties of learning shown in Figure 1 can also be considered as a reason for this ambiguity and abstraction. Our idea of a learning model to reduce the cognitive load for learners to achieve the learning goal is to provide an easy-to-use environment to support learners to reflect their thinking process in their medical services practices. The ontology for patient psychology, medical services, thinking activities and learning activities are incorporated in the environment. And a user-friendly interface for writing case learning materials is provided.

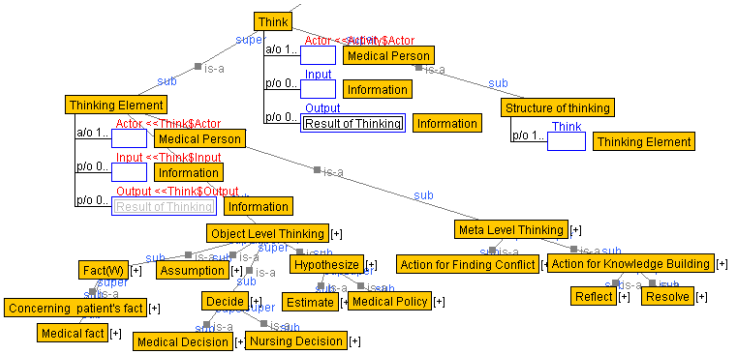


Figure 4. Thinking skill ontology (partial)

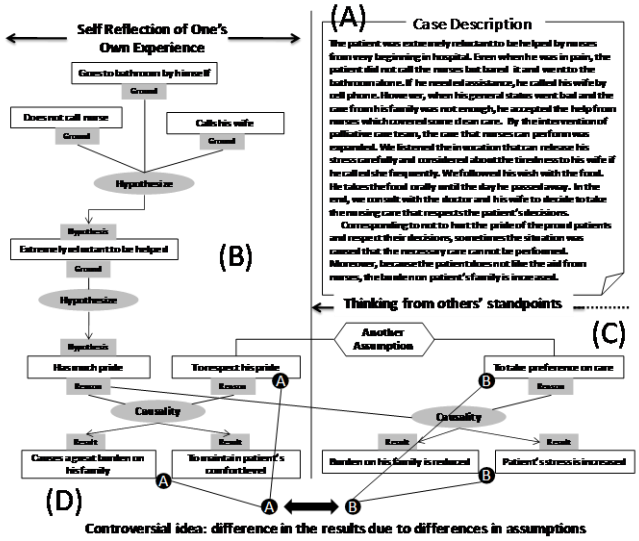


Figure 5. An example of thinking representation in case learning material designing

V. THINKING REPRESENTATION IN CASE DESIGN

Figure 4 shows an overview of a part of the ontology for the thinking process in medical services [8]. Using the concepts in the ontology, the learners externalize the reflection of their thinking process in their experiences in the graphic representation as shown in Figure 4.

Figure 5(A) shows the reflection description of thinking in one's own experience and Figure 5(B) shows its graphic representation. The square nodes represent the assertions and the elliptic nodes represent thinking activities such as "hypothesizing", "finding cause and effect" and so on. Figure 5(C) shows the estimated thinking process of another nurse with a different stance from the learner. Figure 5(D) shows the intended issues (cognitive conflict) to be discussed in the case materials, where a nurse wanted to provide more care, but the patient resisted out of pride, even though this added to the burden of the family in caring for the patient. Meanwhile, she guesses that there may be a nurse who thinks, on the assumption of "care priority", that she should provide more care to the patients even though it may cause strong stress on the patient's mind. Then the learner investigated the advantages and disadvantages of the results of different assumptions.

Associating with the discussion in the previous chapter, (B) the visualization of one's own self-reflection can correspond to the description phase. And (C) according to the assumptions at different standpoints, (D) the discussion set up can correspond to the evocation of knowledge building by cognitive conflicts.

VI. LEARNING ENVIRONMENT FOR THINKING PROCESS TRAINING IN MEDICAL SERVICE EDUCATION

Boud (1985) claims reflection is needed at various points: at the start in anticipation of the experience, during the experience as a way of dealing with the vast array of inputs and coping with the feelings that are generated, and following the experience during the phase of writing and consolidation [13].

Combining the learning strategies based on the goal-attainment model of verbalization (Chapter 3) and the thinking representation in case design (Chapter 4), we developed a learning environment that can conduct the externalization of thinking processes using a model of thinking process for self-dialogue consists of three phases, where the learners are required to be able to conduct high quality thinking for self-dialogue which, to describe high quality reflection on ones' own thinking, to find meaningful conflicts, and to create high quality knowledge in order to overcome the conflicts, by continuously developing their ability using tags.

For the different purposes, we have designed two separate thinking representations for the learning environment. One is the text representation. In medical practice, medical professions are used to writing documents with a similar form of representation, such as the electronic medical records. The other is the graphical representation that provides a learner with an easy-to-reflect overview of the logical structure of the thinking process.

We have developed two thinking support tools, Sizhi and Wuzhi, which correspond to the two representation forms. Moreover, in order to integrate these two forms of representation, we have been developing a bidirectional transformation mechanism between these two representations.

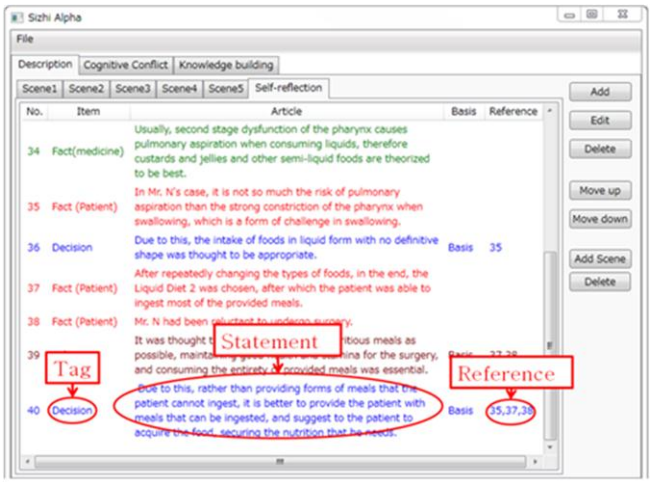


Figure 6. Description phase in Sizhi

Figure 6 shows an example of a case written by a nurse with Sizhi. As shown in the figure, there are three tabs that correspond to the description phase, the cognitive conflict and the knowledge building phase in learning strategies. Each line consists of a statement ID (number), a Sizhi tag, and statement, and may have an additional tag and ID that refer to the logical foundation of the statement in the line. The tags play an important role in encouraging learners to be aware of the logical structure of their own thinking process.

 ISBN: 978-1-61208-181-6

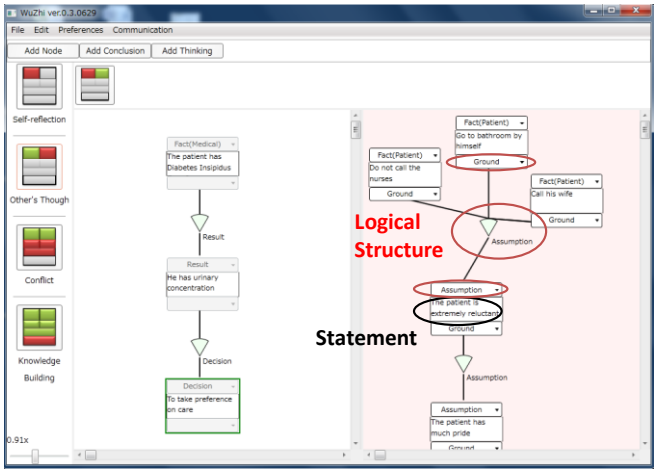


Figure 7. Description phase in Wuzhi

For Sizhi, we will use the thinking ontology mentioned in the above section to clarify the constituents of thoughts, and the learner is required to express the thinking processes using a set of tags as a framework to express the structure of thinking. The set of tags is designed for nurses to reflect on their thinking process for self-dialogue and consists of nine tags: fact (patient), fact (medical), policy/principle, assumption, decision, medical decision, conflicts, reflect and resolve. The nurses' learning task in the case writing is to reflect on their own thinking process in nursing patients and clarify the structure of the thinking process using the tags.

The most important aspect in designing Sizhi is for learners to clearly write their own cases by reflecting on their thinking process using Sizhi tags, and reflect on the thinking process to find meaningful conflicts. To promote learners to gain deep insight into conflicts, for instance, Sizhi encourages learners to find conflicts between the statements with the policy/principle tag, because the policy/principle tag implies the statement is one of logical foundation of the thinking process.

Wuzhi is a learning representation supporting tool that has the same functionalities for supporting the internal dialogue as Sizhi. But the difference between them is that Wuzhi uses a graphical representation at the description phase. For the reason expressed above, the graphical representation can enhance the effect of the descriptions for clarifying their logical structures. Figure 7 shows a medical case written with Wuzhi. Each node in Wuzhi contains the same form of information (tag, statement and reference) as the line in Sizhi.

In summary, for visualizing the invisible, shapeless, complex structure of thinking processes to support knowledge creation, Sizhi provides learners with tags which clarify various thinking processes, and a tab which encourages awareness of the thinking phases, and is designed with the intent to encourage externalization and careful investigation of ideas that follow those processes. Moreover, Wuzhi gives a clear view of representation for description writing and revising. With the help of a thinking ontology, the representation transformation can be conducted smoothly and firmly.

VII. PRELIMIANRY TRIAL USE OF SIZHI

A preliminary experiment was conducted with the help of medical specialists including medial advisors, nurses, medical examination managers, researchers and directors from the Faculty of Medicine, Miyazaki University and the Juntendo University Hospital Group.

Because of the time limitation in the preliminary experiment, the participants could not use Wuzhi. So we focus on the evaluation of the effectiveness of Sizhi. In order to investigate the participants' motivation and their self-evaluation, we conducted two questionnaires, before and after using Sizhi. Based on the analysis of the answers to the questionnaires, we have made the following two findings.

A. Change in Cognition of the Importance of Thinking Skills

We investigated the participants' perception of the importance of thinking skills and the effects it has by self-evaluation. We measured the learners' self-evaluation of how important they think the thinking skills are (perception of importance), and how efficiently they have been using thinking skills before and after the use of Sizhi. To measure the perception of importance, the participants were asked about the target (ex. closely examine whether one's opinions are accurate) which is not related with self-dialogue process, and the distractor (ex. finding flaws in one self), participants were asked to select one from the options: 1- not important at all, 2- not very important, 3- neither, 4- is important, 5- very important. The Figure 8 describes the mean difference in the target column and distractor column before and after using Sizhi. As a result, we found that as the preliminary experiment progressed, the target became higher and the distractor became lower. This result suggests that the understanding of the importance of thinking skills increased by using Sizhi.

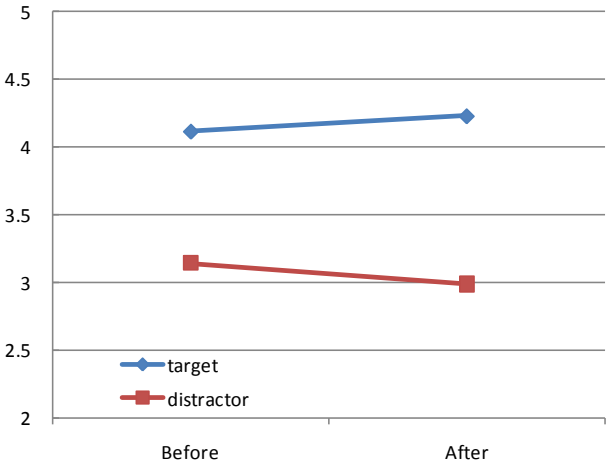


Figure 8. Changes in cognition of the importance of thinking skills before and after using Sizhi

 ISBN: 978-1-61208-181-6

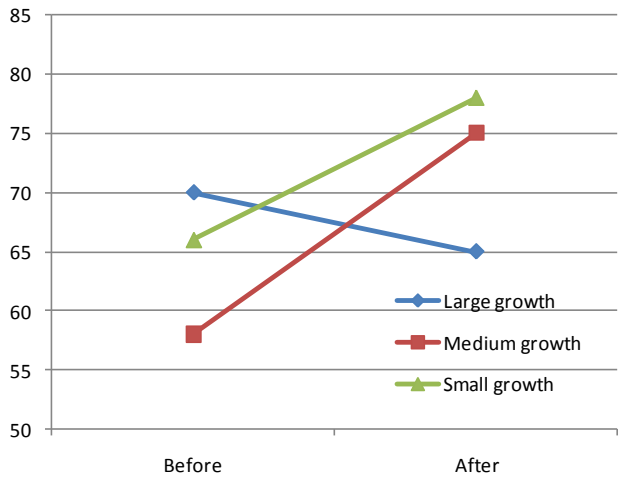


Figure 9. Changes in self-evaluation for each growing before and after using Sizhi

B. *The Change in Self Evaluation of Thinking Ability*

We asked the learners for a self-evaluation of their thinking ability. When we checked the differences in the mean after using Sizhi, the self-evaluation improved. Additionally, we split the learners into 3 groups of 5 (large advance, medium advance, low advance) people according to the magnitude of the change in cognition of importance before and after using Sizhi, and analyzed the self-evaluation (Figure 9). As a result, we saw an improvement in self-evaluation in the medium advance group and low advance group, but we could not see any change in the large advance group. The numbers of subjects were too few for this analysis so we could not conclude any statistically significant differences, but the lack of change in self-evaluation in the large advance group suggests that the bigger the advance, reflecting and evaluating oneself becomes more difficult. To support long term thinking skill mastery, it is important to consider how we can help people like this.

VIII. CONCLUSION

In this paper, we organized a learning model which promotes reflective learning of the case-method for medical service education. As an implementation of the learning model, we have established a learning environment that supports learners to reflect on their thinking process in their experiences by a learning strategy which consists of three case-writing phases: the description phase, the cognitive conflict phase, the knowledge building phase. The final goal of this research is not to make contributions to technological improvement in medical service education but to conduct a proposal of a rational learning model for medical service education. The full educational program we have been developing consists of two parts, that is, one for training thinking skills for internal dialogue and one for training thinking skills for discussion. In this paper, concerning the former part, we have discussed the design rationale of two learning environments, Sizhi and Wuzhi. In our project, as an implementation of the latter part, we have also conducted educational discussion-style workshops at three hospitals. Currently, to shift from face to face discussion to ICT-mediated collaborative learning by integrating, we are developing a learning environment which includes Sizhi and Wuzhi as functional components. In a future paper, we will discuss the design rationale of the learning environment and show evaluation of educational effectiveness through trial use and report on our efforts to put it to practical use in medical service education.

ACKNOWLEDGMENT

Grateful thanks are expressed to Prof. Nobuhiro SATO of Juntendo University, Prof. Kenji ARAKI and Associate Prof. Muneou SUZUKI of Miyazaki University, Dr. Tomohiro NABETA and Dr. Taisuke OGAWA of JAIST for their cooperation.

REFERENCES

[1] Yoshikawa, H. , “Introduction to Service Science”, Journal of the Japanese Society for Artificial Intelligence, vol. 23, No. 6, pp. 714-720, 2008. (In Japanese)

[2] Keio Business School: “Theory and Practice of Case Method”, Toyo Keizai Inc., Tokyo, 1977. (In Japanese)

[3] Hyakkai, S, “Learning by Case Method”, Gakubunsha Inc., Tokyo, 2009. (In Japanese)

[4] Ishida, H., Hoshino, H., and Okubo, T., “Case Book 1: Intorduaction to Case-Method”, Keio University Press, Tokyo, 2007. (In Japanese)

[5] Tomida, E., and Maruno, J., “The Current Study to Arguments as Thinking”, Japanese Psychological Review, vol. 24, No.2, pp. 187-209, 2004. (In Japanese)

[6] Billig, M., “Arguing and thinking: A rhetorical approach to social psychology”, Cambridge University Press, Cambridege, UK, 1987.

[7] Kuhn, D., “The skills of argument”, Cambridge University Press, Cambridge, UK, 1991.

[8] Cui, L. et al., “A Model of Collaborative Learning for Improving The Quality of Medical Services”, Proceedings of The 6th International Conference on Knowledge, Information and Creativity Support Systems, pp.112-121, 2011.

[9] Morita, Y., Cui, L., and Kamiyama, M. “Learning program that makes thinking the outside and presses knowledge collaboration skill development”, The Institute of Electronics, Information and Communication Engineers Technology Research Report, vol. 111, No. 98, pp.7-12. (In Japanese)

[10] Ito, T., “Effects of Verbalization as Learning Strategy: A Review”, Japanese Journal of Educational Psychology, vol. 57, pp. 237-251, 2009. (In Japanese)

[11] Argyris, C. and Schön, D., “Theory in Practice: Increasing Professional Effectiveness,” Jossey-Bass, San Francisco, 1974

[12] Boud, D. and Walker, D., “Experiencing and Learning: Reflection at Work,” Deakin University Press, Geelong, 1991

[13] Boud, D., “Introduction: What is Reflection in Learning,” In: Boud, D., Keogh, R. and Walker, (eds.) Reflection: Turning Experience into Learning, Kogan Page Ltd, London, 1985, pp.7-10.

 ISBN: 978-1-61208-181-6

A Context-Aware Framework for Semantic Indexing of Research Papers

Maryam Tayefeh Mahmoudi[1,2], Fattaneh Taghiyareh[1], Koushyar Rajavi[1], Mohammad Saleh Pirouzi[1]

[1)] School of ECE, College of Engineering
University of Tehran, Iran
[2)] Knowledge Management & E-Organizations Group
IT Research Faculty, Research Institute for ICT (ITRC), Tehran, Iran
Emails: {mahmodi@itrc.ac.ir, ftaghiyar@ut.ac.ir, k.rajavi@ece.ut.ac.ir, s.pirouzi@ut.ac.ir}

***Abstract*—Automatic indexing and annotation of publications have a significant role in retrieving and processing required papers from the massive amount of existing papers in the databases. In this paper, a framework for indexing research papers based on domain ontology is represented. The domain ontology, which is constructed for this purpose, is on agent science and technology. The initial step to indexing is to recognize the major concerns and the basic constituents in the title of papers, which has been accomplished through proposing a few NLP-based rules. To annotate each paper, the mentioned ontology and WordNet are employed. Experimental results on about 155 research papers lead us to estimate that our framework is capable of semantic indexing in about 80 percent of the situations. Since we have considered the ontology separately from the constituents of the whole system, the proposed framework is domain-independent and can be applied to any other domain ontology.**

Keywords-Indexing; domain ontology; research paper; WordNet; incremental learning.

I. INTRODUCTION

Semantic Indexing has an influential role in managing tremendous amount of publications in databases. The goal of semantic indexing is to offer more effective search and categorization services. There exist various methods for this purpose such as Latent semantic indexing (LSI) and concept indexing (CI), which are among information retrieval techniques [1, 2]. Although they have empirical success, they suffer from the lack of interpretation for the low-rank approximation and, consequently, the lack of controls for accomplishing specific tasks in information retrieval [3]. To overcome the existing deficiencies, there is a tendency to more potential schemes like ontology and semantic web. Domain ontology seems to be an appropriate tool for supporting indexing methods which can be widely used for knowledge and content processing applications [4, 5, 6]. In the meantime, the construction of domain ontology relies on domain modelers and knowledge engineers that are typically overwhelmed by the potential size, complexity and dynamicity of a specific domain [7]. To overcome the barrier of constructing exhaustive domain ontology, annotating or indexing may again be an appropriate alternative to enable the ontology with the potential of learning [8]. Thus, close examination of the issue reveals that indexing plays a major role both in publications' storage management and consequently incremental ontology learning.

Taking the rapid growth in the number of research papers into account, turns into deployment of appropriate indexing method for facilitating both storage and retrieval purposes [9].

In this paper, we propose a context-aware framework for semantically indexing research papers based on domain ontology and NLP-based rules. The domain ontology which is constructed for this purpose is on agent science and technology issue, while the NLP-based rules are achieved through processing huge amount of research papers' titles. The main reason that titles of publications are considered as the basis of indexing is that they are informative enough that there is no need to process the whole text. Determining major concern and basic constituent behind the title leads into semantically indexing each paper. By major concern, we mean the major objective and concern of the title that illustrates why and for what reason it is under consideration, while by basic constituent we mean how to realize the major concern, by applying special tools or means [10, 11]. Having a review on existing approaches using major concern and basic constituent reveals the potential of these concepts in indicating the main objective behind the whole text [12, 13]. To automate the above mentioned process, some NLP-based rules are proposed. It is no doubt that by matching the major concern and basic constituent with the existing concepts in the domain ontology, annotation will be accomplished. It is to be added that WordNet can also be a supportive tool for finding the closest concepts in an unmatched cases.

The rest of the paper is organized as follows: Section 2 reviews some of the previous works which have been done in the area of indexing and annotating. Section 3 describes our suggested framework. In Section 4, experimental results are analyzed and Section 5 sketches out the conclusion and future works.

 ISBN: 978-1-61208-181-6

II. RELATED WORK

Due to the rapid growth in number of publications, organizing papers and documents in a certain database has become more important than before, especially for store, search and retrieval purposes [14].

In the meantime, there exist various indexing or annotating methods which are discussed as follows:

Some focus on phrase-based document similarity via index graph model. This method has the potential of detecting any-length phrase match from the current document to all the previously seen documents in the data set by just scanning it and extracting the matching phrases from the document index graph [15]. Index-Filter is another method, which uses indexes built over the document tags to avoid processing large portions of the input document [16]. In addition to data structure for indexing XML documents based on relative region coordinates which describe the location of content data in XML documents is also mentionable [17]. With respect to managing the large spatial ontologies, spatial index for improving the efficiency of the spatial queries are deployed [18].

In addition to the methods discussed above, Latent Semantic Indexing (LSI) or Latent Semantic Analysis (LSA) and also concept indexing (CI) are mentionable. Those methods improve the detection of relevant documents on the basis of terms found in queries [19]. The most challenges to LSI focused on scalability and performance. LSI requires relatively high computational performance and memory in comparison to other information retrieval techniques [20].

PubSearch uses a citation based retrieval system [14] which generates a web citation database from online scientific publications that are available over the internet. Random indexing is another method which is based on an incremental word space model [21]. The basic idea of Random Indexing is to accumulate context vectors based on the occurrence of words in documents.

It is not to be disregarded that ontologies have significant role in semantic annotation, too [5, 22]. They are being widely used in information retrieval (IR) either for performing semantic indexing of documents or to produce a better organization of retrieved documents [23]. In this respect, document indexation methods, specifically in large-scale web search engines, support the retrieval of documents that might contain some parts related to the query [24]. Linguistic annotation is also an important field in natural language processing that involves classification of text into a predefined set of values [25]. Improving the semantic capability of ontology-based indexing method by major concern and basic constituent is our concern in this paper.

III. SUGGESTED FRAMEWORK

A. The overall Structure of Proposed Framework

As it has been mentioned before, in large-scale databases of research papers, applying a well–defined indexing method plays a significant role to retrieve desired papers. In this respect, we propose a context-aware framework that seems to be capable enough to facilitate such a process. Figure 1 illustrates the details of proposed framework.

As it is illustrated in Figure 1, each time that a paper is uploaded to the database, it is necessary to be indexed. For this purpose, extracting the title of the paper and parsing it is required in order to find major concern and basic constituent.

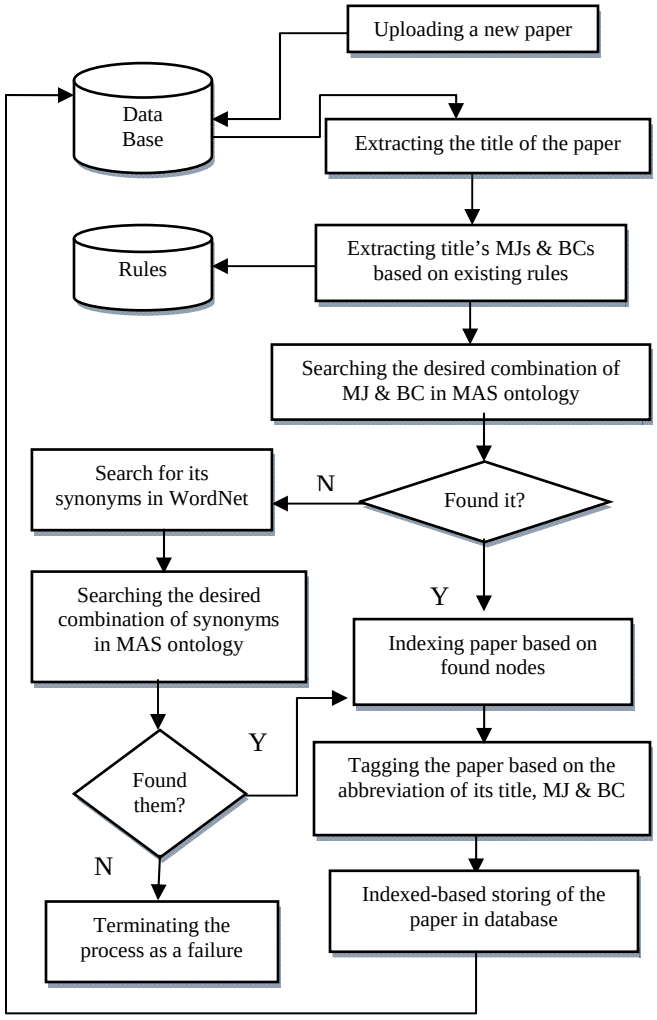


Figure 1. Flowchart of proposed indexing framework

Defining suitable rules as well as domain ontology can realize the indexing process appropriately. For this purpose, we are employing the ontology of agent science and technology that we have developed in Multi Agent Science Laboratory of University of Tehran for this purpose.

Searching the desired combination of major concern and basic constituent in agent science and technology ontology and finding the related nodes reveals the appropriate keywords for tagging and indexing the paper. It is to be noted that in cases where the major concern and basic constituent do not match any node of ontology, WordNet

seems to be a good realm for substituting alternative words. If the related words, in any of the mentioned ways be found, the paper will be tagged, indexed and respectively stored in the database. Otherwise, the process will be terminated as a failure. In the following sections, we briefly discuss each major constituent of proposed framework illustrated in Figure 1 including the functionality of major concern and basic constituent in indexing process. Followed by explaining the role of ontology and WordNet to support appropriate semantically indexing of papers and respectively storage of them.

B. The proposed indexing method based on major concerns and basic constituents

As previously mentioned, each title usually has two parts; major concern (MJ) and basic constituent (BC). Major concern is the part which explains about the main objective of the paper, while Basic Constituent mainly focuses on the methods, techniques or tools which were used to reach the objective in major concern [10, 12].

Reviewing several numbers of titles lead us to the following structure for MJ and BC. Four main parts are considered for MJ as follows:

1) Action part; which is mainly a verb.
2) Direct object; which is a noun or a pronoun that receives the action of a verb or shows the result of the action. It answers the question "What?" or "Whom?" after an action verb.
3) Indirect object; which is the recipient of the direct object and answers the question "To whom?" or "For whom?". It usually follows a preposition
4) Adverb/ Adjective part; which can modify verbs, adjectives, clauses, sentences, and other adverbs. It typically answers questions such as "how?", "in what way?", "when?", "where?", and "to what extent".

It has to be mentioned that some conjunctions like "in" and "for" followed by a verb usually yield into having two layers for MJ, which follow the same structure as mentioned above. Obviously, one or some of these parts may be absent in a title.

For BC, most of the time, maximum of one layer seems to be sufficient.

Having studied several titles, we gathered some rules, which were used to extract MJ and BC from a title. Certain conjunctions and prepositions can be signs of MJs and BCs. We prioritized some prepositions and conjunctions over others. Table 1 is a list of some of these prepositions and conjunctions.

TABLE I. LIST OF PRIORITIZED PREPOSITIONS AND CONJUNCTIONS

Priority	Preposition/ Conjunction
PR1 (BC)	Based on, on the basis of, on the ground of, using, making use of, taking into, ….
PR2 (MJ)	With the purpose of, with the aim of, in order to, in order that, …
PR3 (BC, MJ)	with, by, in, for, via, at, from, about, across, after, against, along, among, around, before, behind, beside, during, inside, instead of, onto, outside, over, since, through, under, within, …
PR4	and, of, into, like, without, both, together with, as, neither, either, as well as, rather than, than, …

Using the prepositions and conjunctions in the table above, we are able to detect BC and MJ before or after these words.

Employing a NLP parser can facilitate this process. For example, consider the title "Extending process automation systems with multi-agent techniques", as it is illustrated in the table, basic constituent can be found after "with", while before "with" we have major concern. It is to be noted that in MJ part, "extending" plays the role of action while the "process automation system" refers to direct object. Processing some complicated titles, necessitate more rules. For instance, "An agent-based signal processing in node environment for real-time human activity monitoring based on wireless body sensor networks" is a complicated title including several conjunctions and prepositions. Figure 2 illustrates MJ and BC of the title in detail.

MJ

	Action-part	Adverb-part	Direct Obj.	Indirect Obj.
MJ_1	**processing**	**agent-based**	**signal**	**Node environment**
MJ_2	**monitoring**	**Real time**	**Human activity**	-

BC

1st Layer :	2nd Layer
wireless body sensor networks	-

Figure 2. An example of major concern & basic constituent

Having reviewed large amount of titles, yield several rules for distinguishing MJs and BCs, Such as:

- If we find “via” or “based on” in the title, the following word or phrase will be BC.
- If we find “for” in the title, with a verb following, the rest of the title will be accounted as the 2nd layer of MJ.

C. *Ontology Processing*

After extracting MJ and BC based on the rules discussed in the previous section, combinations of MJ and BC are applied for seeking in the domain ontology. If the corresponding node is found, the keyword for annotating is achieved; otherwise closely related words or synonyms from WordNet have to be extracted for the same purpose. In this manner, indexing process based on the active nodes of ontology is realized. Vise versa, in the cases where not any related node is found, the process will be terminated and a failure notice will be issued. In situations where hierarchical ontology learning is considered, the new concept will be added in to the closest node of ontology. This would be our future trend of research in this subject.

Our domain ontology contains 200 nodes, with the depth of eight, describing agent science and technology. Figure 3 reveals a part of that ontology.

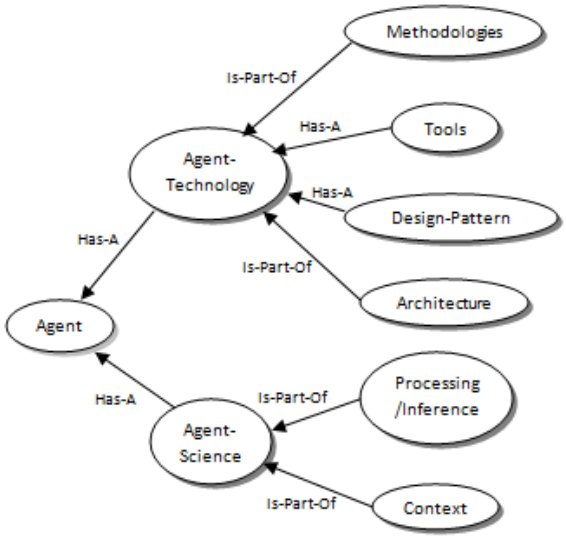


Figure 3. Part of the agent science & technology

IV. EXPERIMENTAL RESULTS

In order to evaluate the proposed approach, a data set of about 155 research papers in domain of agent science and technology which are collected from different conferences is considered.

We have indexed the research papers using the proposed framework. Figure 4 shows the pseudo code of the proposed framework.

Table 2 represents some examples of what our system produced for MJ, BC and related nodes in ontology. The papers which are shown in Table2 are as follows:

1: Group Communication based approach for Reliable Mobile Agent in information Retrieval Applications
2: Using a Dynamic Swarm of Intelligent Agents for Advising Farmers-AgroAgent
3: An Intelligent Inter Database Retrieval System Based on Multi-agent
4: The Personalized Information System of Lib2.0 Based on Agent

```
Extract Title;
  Parse (Title);
    Extract (MJ, BC);
      Search in Ontology (MJ, BC);
        Mark Related Nodes in Ontology;
          If (Marked Nodes == empty)
        Find Synonyms (MJ, BC);
      Search Synonyms in Ontology;
    Mark Related Nodes in Ontology;
Indexed-based Storage;
```

Figure 4. Psuedo code for proposed indexing framework

TABLE II. MJ, BC & ACTIVATED NODES IN ONTOLOGY FOR SOME PAPERS

Title	Major Concern (1st layer)	Major Concern (2st layer)	Basic Constituent	Activated nodes in Ontology
1	Group Communication based approach	Reliable Mobile Agent in information Retrieval Applications	-	Application, Mobile, Communication
2	Advising Farmers-Agro Agent	-	a Dynamic Swarm of Intelligent Agents	Dynamic, Reasoning
3	An Intelligent Inter Database Retrieval System	-	Multi-agent	Reasoning, Agent
4	The Personalized Information System of Lib2.0	-	Agent	-

As it is shown in table 2, our system has failed in finding an appropriate node for the 4th title because it wasn't able to find any of the words (and their synonyms) in our ontology. In the first two titles, our system has done very good in detecting MJ, BC and also in activating related nodes in ontology. For the 3rd title, despite having correctly extracting MJ and BC, our system mistakenly activated "Reasoning" node because of its similarity to the word "intelligent" which was in MJ of the title. For this reason, in future works, we have to apply better rules to avoid these mistakes.

Our system also shows great ability in finding MJ and BC of second layer. For example, for the title "Scalability and Load Balancing for Multiplatform Communication System Architecture based on Intelligent Agents", it gives "scalability and load balancing" as the MJ of the first layer and "Multiplatform Communication System Architecture" as the second layer of MJ, and therefore is able to detect "scalability, communication, architecture, reasoning" as the

 ISBN: 978-1-61208-181-6

corresponding nodes in the ontology. Overall, we were satisfied by the results our framework produced in processing MJ and BC for about 130 papers of the 155 papers.

In this paper, we used precision and recall measurements to judge the efficiency of our method. The "Precision" is calculated as the proportion of relevant retrieved documents to the number of retrieved documents and "Recall" is defined as the proportion of relevant retrieved documents to total number of relevant documents [26, 27].

$$\text{Precision} = TP/(TP + FP) \quad (1)$$
$$\text{Recall} = TP/(TP+FN)$$

where TP, TN, FP and FN denote true positives, true negatives, false positives, and false negatives, respectively.

Experimental results reveal that our system was able to index 118 number of papers correctly, while for 9 papers, our system issued a failure. It also made mistake in indexing 28 of the papers. Therefore, as it is illustrated in Figure 5, the

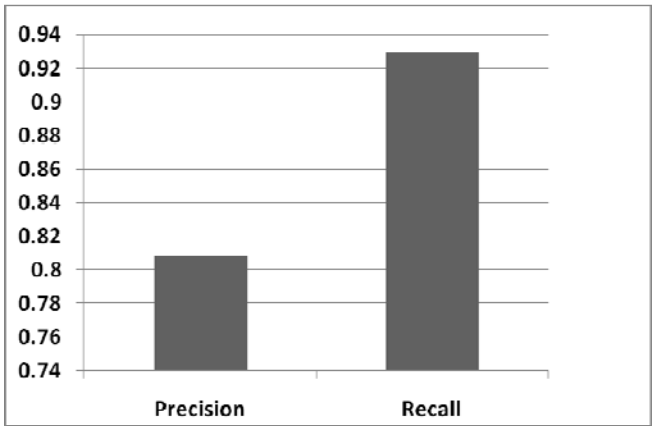


Figure 5. Precision and recall of experimental results

precision ratio is equal to 118/(118+28)= 0.81, while the recall ratio is 118/(118+9)=0.93.

In essence, our framework was able to correctly index 3/4 of the papers. The failures were mainly because of the following reasons: 1) Incompleteness of our ontology. 2) Lack of rules for extracting MJ and BC. 3) WordNet's inability to find scientific and agent-related words and phrases, and therefore not finding their synonyms. These problems can easily be resolved in future and as a result, the efficiency and correctness of our framework will be improved.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed an ontology-based indexing framework that can be used for managing the storage of papers in database. The proposed approach can also be an appropriate method for investigating the incremental learning of ontology. Distinguishing major concern and basic constituent of the title is our mean to find the corresponding nodes of ontology. To implement the proposed approach some NLP-based rules are proposed. It is mentioned in the paper that, in cases where there is no corresponding node in the ontology, WordNet is an appropriate supportive tool. Improving the rules from one perspective, and applying the proposed method for incremental ontology learning from another perspective are our future researches in this issue.

REFERENCES

[1] T. Hofmann, "Probabilistic latent semantic indexing," Proc. of the 22nd annual Intl. ACM SIGIR conference on Research and development in information retrieval, 1999, pp. 50-57.

[2] J. M. Gómez, J. C. Cortizo, E. Puertas, and M. Ruiz, "Concept Indexing for Automated Text Categorization," Natural Language Processing and Information Systems, Lecture Notes in Computer Science, Vol.136, 2004, pp. 495-502.

[3] J. Dobsa, " Comparison of information retrieval techniques: Latent semantic indexing (LSI) and Concept indexing (CI)," Solomonovi seminarji, Faculty of Organization and Informatics, Varazdin, University of Zagreb, Feb. 25, 2007.

[4] T. B. Lee, J. Hendler, and O. Lassila, "The semantic web," Scientific American, Vol. 284, No. 5, 2001, pp. 35–43.

[5] J. Kohler, S. Philippi, M. Specht, and A. Ruegg, "Ontology based text indexing and querying for the semantic web," Knowledge-Based Systems, Vol. 19, 2006, pp. 744–754.

[6] J. J. Chen and R. Sison, "Domain ontology learning," 5th Philippine Computing Science Congress, 2005, pp. 161-165.

[7] D. Sánchez, "Domain ontology learning from the web," The Knowledge Engineering Review, Vol. 24, Issue 4, 2009, pp. 413-413.

[8] M. Hazman, S. R. El-Beltagy, and A. Rafea, "A survey of ontology learning approaches," Intl. Journal of Computer Applications,Vol. 22, No.9, May 2011, pp. 36-43.

[9] L. Feigenbaum, M. N. Roy, and B. H. Szekely, W.C.Yung, "Ontology based text indexing," United States Patent Application 20080288442.

[10] K.Badie & M.T.Mahmoudi, " An approach to generating new ideas based on linking frames of concepts," ECCBR'04, Computational creativity workshop, 7th European Conf. In Case-based reasoning, Madrid, Spain, Aug 30- Sep 2, 2004.

[11] M.T.Mahmoudi & K.Badie, "Content determination for composite concepts via combining attributes' values of individual frames", IKE'04, Intl. Conf. On Information & Knowledge Engineering, Las Vegas, USA, Jun 21-24, 2004.

[12] K.Badie. M.T.Mahmoudi, "A Computational Framework for Manipulating an Issue from the View-Point of Other Issues," 14th Intl. Cong. of Cybernetics and Systems of WOSC – ICCS'08 ,Wroclaw, Poland, 2008, pp. 9 – 12.

[13] K.Badie, M.T.Mahmoudi, "View-Point Oriented Manipulation of Concepts: A Matching Perspective", IEEE Second International Conference on the Digital Society, ICDS 2008, Sainte Luce, Martinique, February 10-15, 2008.

[14] Y. He, S. C. Hui, and A. C. M. Fong, "Mining a web citation database for document clustering," Applied Artificial Intelligence Journal, Vol.16, 2002, pp. 283-302.

[15] K. M. Harnmouda and M. S. Kamel, "Phrase-based document similarity based on an index graph model," Second IEEE Intl. Conf. on Data Mining (ICDM'02), 2002, pp. 203-210.

[16] N. Bruno, L. Gravano, N. Koudas, D. Srivastava, "Navigation vs. index-based XML multi-query processing," 19th International Conference on Data Engineering (ICDE'03), 2003, pp. 139-150.

 ISBN: 978-1-61208-181-6

[17] D. D. Kha, M. Yoshikawa, and S. Uemura, "An XML indexing structure with relative region coordinate," 17th Intl. Conf. on Data Engineering, 2001, pp. 313-320.

[18] E. Dellis and G. Paliouras, "Management of large spatial ontology bases," Workshop on Ontologies-based techniques for DataBases and Information Systems (ODBIS) of the 32nd Intl. Conf. on Very Large Data Bases (VLDB 2006), 2006, pp. 102-118.

[19] S. T. Dumais, T. K. Landauer, and M. L. Littman, "Automatic cross-linguistic information retrieval using Latent Semantic Indexing," SIGIR'96, Workshop on Cross-Linguistic Information Retrieval, 1996, pp. 16-23.

[20] G. Karypis and E. Han, "Fast Supervised Dimensionality Reduction Algorithm with Applications to Document Categorization and Retrieval," 9th International Conference on Information and Knowledge Management (CIKM), 2000, pp. 12 – 19.

[21] M. Wan, A. Jonsson, C. Wang, L. Li, Y. Yang, "A random indexing approach for web user clustering and web prefetching," 2011 Workshop on Behavior Informatics, Shenzhen, China, 2011.

[22] N. Sugiura, K. Masaki, F. Naoki, I. Noriaki, and Y. Takahira, "A domain ontology engineering tool with general ontologies and text corpus," 2nd Workshop on Evaluation of Ontology based Tools, 2003.

[23] A. J. Yepes, R. B. Llavori, and D. R. Schuhmann, "Ontology refinement for improved information retrieval," Information Processing and Management, Vol. 46, 2010, pp. 426–435

[24] P. Martin and P. Eklund, "Embedding knowledge in Web documents," Computer Networks, Vol. 31, 1999, pp. 1403–1419.

[25] W. W. Chapman and J. N. Dowling, "Inductive creation of an annotation schema for manually indexing clinical conditions from emergency department reports," Journal of Biomedical Informatics, Vol. 39, 2006, pp. 196–208.

[26] D. L. Olson and D. Delen, "Advanced data mining techniques," Springer, 1st edition, February 1, 2008, pp. 138-138.

[27] C. D. Manning, P. Raghavan, and H. Schütze, "Introduction to information retrieval," Cambridge University Press., 2008.

 ISBN: 978-1-61208-181-6

Taste It! Try It! – A Semantic Web Mobile Review Application

Monika Kaczmarek, Agata Filipowska, Jakub Dzikowski, Szymon Łazaruk, and Witold Abramowicz
Department of Information Systems, Faculty of Informatics and Electronic Commerce, Poznan University of Economics, Poznań, Poland
{ m.kaczmarek; a.filipowska; j.dzikowski; s.lazaruk; w.abramowicz}@kie.ue.poznan.pl

***Abstract*—Web-based review systems provide a valuable service to consumers allowing them to share opinions on various goods and services. In order to improve the comparability of opinions as well as their retrieval and further application, systems more and more often take advantage of semantic annotations and Semantic Web technologies. However, as users are not really interested in delivering semantic annotations of content, specific tools incorporating incentives' mechanisms are needed to transform the syntactic content into the machine understandable one. This paper presents a Semantic Web mobile review application, supporting semantic-based user profiling and contextual semantic search, benefiting from linked data and equipped with Web 2.0 motivations mechanisms.**

Keywords – Semantic Web; semantic annotation; Linked Open Data; incentives; Web 2.0; review systems.

1. INTRODUCTION AND MOTIVATION

There exists an enormous amount of reviews on goods and services published by users on the Web every day. Reviews made by consumers, shared opinions and experience, have become an important source of valuable information that can be used by recommendation systems. However, as many consumers prefer to use free text to express their opinions, the difficulty in structuring these reviews using information extraction techniques [3] makes opinions' selection and retrieval processes as well as utilization of retrieved opinions not accurate enough [15].

Therefore, in order to improve the accuracy of mentioned processes, review systems more and more often take advantage of semantic annotations and Semantic Web technologies [6]. However, the interest of users to contribute to the creation of the semantic content is rather low, due to [8][9]:

- rather high barrier of entry - creation of semantic annotations requires specific skills and expertise in the domains such as ontologies, logic and knowledge representation;
- lack of incentives - most of the semantic applications are difficult to use and lack built-in incentives inducing users to use them;
- lack of clear benefits - the benefits of using semantic content are in many applications decoupled from the effort of creating the semantic content.

Thus, in order to address the problem of data structuring and engage users in the process of creating semantic annotations, tools need to incorporate adequate incentives, e.g., benefiting from the Web 2.0 paradigm [17].

In addition, the Web has evolved "from a global information space of linked documents to one where both documents and data are linked" [12]. This evolution is supported by a set of best practices for publishing and connecting structured data on the Web known as Linked Data. The most visible example of adoption and application of the Linked Data principles is the Linking Open Data project [13] aiming at identifying existing data sets that are available under open licenses, converting them to RDF [14] according to the Linked Data principles, and publishing them on the Web (as Linked Open Data (LOD)). Thus, instead of developing new standalone ontologies to be used within review systems, it is desirable to take advantage of the LOD paradigm. The machine-readable data coming from various data sets with explicitly defined meaning can provide better access to various information sources (by supporting comprehensive answers to queries over aggregated data), thus, leading to enhanced user experience.

In this paper we present a Semantic Web mobile application for reviewing objects supporting semantic-based user profiling and contextual semantic search, benefiting from LOD sets and equipped with Web 2.0 motivations mechanisms. The application is being developed within the INSEMTIVES project [10] focusing on mechanisms motivating users to dedicate more time and resources in the participation in the process of semantic content creation.

The goal of the developed application is to make annotating process sufficiently easy for end-users' acceptance while providing added value through the ease of integrating data and reasoning on it. The work conducted encompassed both the research and practical related aspects. On the one hand, the aim was to contribute to a general understanding of the problem and on the other hand, the aim was to develop a system that could not only be used as a proof for testing, but also could constitute a fully fledged tool to be used by users. Thus, the System Development Method (SDM) was utilized [11] that "allows the exploration of the interplay between theory and practice, advancing the practice, while also offering new insights into theoretical concepts" [11]. The approach

 ISBN: 978-1-61208-181-6

followed consisted of three main steps. First, the concept building phase took place, which resulted in the theoretical concepts presented in this paper. The next step is the system building encompassing development of a system based on the theoretical concepts established. The last step will be the system enhancement and evaluation.

The paper is organised as follows. Next section presents shortly the related work. Within the following section, the general framework of the proposed solution is presented including the vision of the tool, approach to the semantic annotation followed, supported functionalities and the motivations mechanisms applied. Finally, the paper concludes with final remarks.

2. RELATED WORK

Recommendation systems attempt to predict items a user may be interested in, given some information about user's preferences and past behaviour i.e., a user profile [1][15]. Most existing recommender systems take advantage either of:

- collaborative filtering techniques i.e., analyzing past actions and behaviour of all users in order to identify interesting associations between them or between the objects, which can be used to make recommendations to a single person (memory-based collaborative filtering (e.g., [18]) and model-based collaborative filtering (e.g., [19]));
- content-based methods i.e., recommending objects by analyzing the associations between user's past choices and descriptions of new objects [2][16] or
- hybrid filtering methods combining two previous ones.

A typical recommendation mechanism analyzes the user context (a user profile, if available), and presents to the user one or more descriptions of objects that may be of their interest. Recommendation mechanisms may be used in pull (recommendations are explicitly requested) or push mode (recommendations are made when a user did not ask for them). In either way, the recommendation should be personalized [20]. Following [16], different levels of personalization can be distinguished starting from coarse grained ones (e.g., relying on the country of residence) to fine-grained (e.g., based on the recent search history). The process of personalization is accurate, if the system possesses accurate information on a user as well as the object/topic the user is interested in, and the information is machine-understandable.

As online reviews are increasingly becoming the de-facto standard for measuring the quality of various goods and services, their sheer volume is rapidly increasing so that processing and extracting all meaningful information manually in order to make an educated purchase becomes impossible [32]. As a result, there has been a trend towards systems automatically summarizing opinions from a set of reviews and displaying them in an easy to process manner [32][33][34]. One of the approaches followed is the aspect-based sentiment summarization taking as an input a set of users' reviews for specific goods and producing a set of relevant aspects, an aggregated score for each aspect, and supporting textual evidence [32]. The quality of this aspect-based summarization may be highly increased, if an a priori knowledge domain and the labelling of the portal are also considered [32].

Although much has been done, the current generation of recommender systems still requires further improvements including methods representing user behaviour, incorporation of various contextual information into the recommendation process, utilization of multi-criteria ratings [1] as well as extracting information from free-text comments left by customers [3]. Although much progress has been made in the area of tools automatically producing structured reviews from unstructured text [3][32][33][34], human involvement is still required. Therefore, the application of semantics and the idea to apply appropriate incentives to encourage people to create semantic annotations should be considered [9].

According to [5], semantics is one of top ten most promising technologies of the future. The Semantic Web paradigm constitutes a major step in the evolution of the Web. It is to enable machines to understand the meaning of information on WWW via extending the network of hyperlinked human-readable web pages by inserting machine-readable metadata about the Web content and information on how they are related to each other, thus, enabling automated reasoning [6]. Its main goal is to make the Web content not only machine-readable, but also understandable by using semantic annotations. A semantic annotation is machine understandable, if it is explicit, formal, and unambiguous and this goal is usually reached by using ontologies [7]. Semantic review systems are those whose performance is based on some knowledge base defined as e.g., ontology [1][4]. The application of ontology within the review system: semantically extends descriptions of user opinions; allows to complete the incomplete information through inferences; semantically extends descriptions of user contextual factors; allows for the dynamic contextualization of user preferences and opinions in specific domains; guarantees the interoperability of system resources and the homogeneity of the representation of information; improves communication processes between agents and between agents and users [4].

As already mentioned, the Web has also evolved into the Web of Data [12] by using a set of best practices for publishing and connecting structured data on the Web known as Linked Data. The Linking Open Data project [13] identifies existing data sets that are available under open licenses, converts these to RDF [14], and publishes them on the Web. The examples of datasets encompass well-known DBPedia, Geonames or Freebase. The content of the Linked Data cloud is diverse in nature, comprising e.g., [12]: data about geographic locations, people, companies, books [21], scientific publications, movies, music, television and radio programmes, genes, proteins, drugs and clinical trials, online communities, statistical data, census results, and reviews (Revyu system [22]).

Currently, because of the rising popularity of Web 2.0 tools, product review forums have become ubiquitous and

 ISBN: 978-1-61208-181-6

more and more websites provide platforms and tools for customers allowing them to share with others their personal evaluations and opinions on products and services, e.g., Yelp, Goodrec Urban or Tripadvisor. These systems provide large amount of reviews and offer recommendations on goods and services. Although quite successful, the precision of browsing and searching the reviews previously submitted is far from being perfect and usually no summary of existing reviews is provided. However, there exists few solutions incorporating semantic technologies and therefore providing more precise search results e.g., [23][24][3]. However, they suffer from a lack of user-generated content. The most promoted semantic recommender system seems to be the Revyu system – a generic reviewing site based on the Linked Data principles and the Semantic Web technology stack. Although the tool itself is worth noting, it also lacks on focus (a user may review anything he wants) and on incentives mechanisms encouraging users to provide semantic content of high quality..

The social phenomenon of Web 2.0 is well recognised in the literature and well visible in the everyday life [17]. With mobile-multimedia devices capable of continuum data transmission, social interactions on the Web are to be moved to the new level as big numbers of different individuals who wish to contribute towards some joint project or community may be easily linked together [25][26]. This trend is also visible when it comes to social tagging sites like Flickr and Delicious or already mentioned various recommendation sites. People sometimes work for free, motivated either out of intrinsic enjoyment [27], social reward [28] or by using financial compensation (e.g., Mechanical Turk). The key for success of every user-contribution based system is the incentives mechanism applied. The gratification system should be as attractive as possible and each award should motivate a user towards further contribution. The success of Farmville on Facebook showed the power of funny badges and medals published on the Facebook wall. When it comes to review systems, some applications use simple flat points to award users for their contribution (e.g., Gastronauci), some use complex system of badges (e.g., Foursquare) or stamps (e.g., Gowalla). In addition, some of them offer publishing information on user activity on the Facebook wall (e.g., Urban spoon, mygoodeats), which additionally motivates users and is a great way of attracting new users to sing up.

Taste it! Try it! is to provide the following additional value in comparison to the currently existing solutions: structuring and disambiguation of the reviews by using domain knowledge, complex ontology-based description of objects integrated with the LOD cloud; semantic-based user profiling and personalization of search results; incentives to contribute to the system following appropriate usability and social design guidelines.

3. TASTE IT! TRY IT! APPLICATION

The Taste it! Try It! application is targeted at two groups of end-users: data producers (contributors) - providing reviews of places, and data consumers (beneficiaries) - interested in the content produced by the application, i.e., looking for opinions on various places.

3.1 STORYBOARD

The Taste It! Try It! Application supports the creation of semantically annotated reviews using mobile devices in a user-friendly manner. The storyboard supported by the system is as follows. A user goes to a restaurant. While being at the restaurant, the user decides to share his opinion on the restaurant and its quality of service factors with other members of the community. He uses Taste It! Try It! to express this opinion. The application starts from capturing the position of the place (using the GPS system in a mobile device). This enables associating the semantically annotated review that is created afterwards with a specific point in space. Then, the user creates a review by providing values to selected features suggested by the application. Additionally, the user may create a free-text comment regarding the object being reviewed. The review is then uploaded to a Taste it! Try it! server and in the background the semantic representation is created. Based on the quantity and quality of created annotations, the user may be awarded with a special title e.g., Polish-cuisine expert, International-food expert. This title is visible to his friends at the community portal, in our example the Facebook portal, with which the application is integrated. In addition, based on the user behaviour and data made available by the Facebook portal, the user profile is created, which is then used in the personalization process. As data acquired from Facebook and other sources is structured, it can be directly mapped to the ontology used by the application. The created annotations are then further on used by a semantic-based recommender system while searching for restaurants fulfilling certain criteria, e.g., vegetarian, low budget, and high quality, in the neighbourhood of a user. As the semantically annotated reviews are linked to LOD sets [12], some more sophisticated reasoning over the data is to be possible and extends the possibilities offered by the system.

Thus, the application is to fulfil the following goals:

- provide semantically-enabled reviews that are sufficiently easy to create for end-user acceptance - the process of attaching the machine understandable semantics should be invisible to the end user;
- keep a user entertained - integration of the proposed application with the social portal such as Facebook and badges, are some of the incentives that are utilised to make the system more attractive to users.
- offer the personalized, semantic, context-aware recommendation process (both push and pull).

3.2 SEMANTIC ANNOTATION AND SUPPORTED FUNCTIONALITIES

Within our work, we followed a hybrid approach to the review creation and in consequence, also to a semantic annotation process. Firstly, we decided to include into our model a feature-based review relying on labelling. Thus, the domain knowledge was utilised in order to identify the

 ISBN: 978-1-61208-181-6

most important dimensions of reviewed objects users may be interested in, e.g., details of the place, food (the quality of served food, how the food tastes, and comments about specific dishes, items or selection); service (such as mainly politeness and timeliness of order delivery); atmosphere (information on the venue such as: decoration, parking, cleanliness, music, etc.), value related (e.g., the quality of goods in comparison to their price).

Our dimensions encompass both the quantitative and qualitative information at the feature/aspect level. The quantitative ones are appropriately aggregated and translated into a "star rating" between one and five stars. The qualitative ones are represented by a set of possible values representing the key sentiments that may be used to express the value of the aspects. Thus, while delivering the review, besides providing specific rates of a restaurant in the various categories, users may specify the restaurant's cuisine, available entertainment, payment options, Internet-access possibility, etc., by selecting the appropriate values from the lists.

All information is expressed in a formal semantic manner in the background. The specific dimensions are linked to the internal ontology developed within the project, with the central concept review. The mentioned ontology provides flexible categorization scheme that assist in further integration of reviews (to produce a consistent description of an object e.g. restaurant) as well as recommendation and search. In addition, the specific concepts in the ontology are linked to the LOD cloud; e.g., the dimension object city is linked to concept city from GeoNames and local is linked to the restaurant concept from DBpedia. Each review made by a user produces additional RDF triples that may be published in the LOD cloud (e.g., 'Quality Restaurant' is located in Paris).

The second approach to the semantic annotation, allows users to introduce into the application the free-text comments regarding any selected aspect of an object being reviewed. While the user is typing the comments, the system on the fly checks the words used and tries to disambiguate them and link them to the existing concepts in the LOD. This is done on the mobile device, if a user is online or at the Facebook portal in other case. In this way additional RDF statements are created – annotating restaurants and their different aspects selected by a user. Free text comment is disambiguated using Wordnet or DBPedia and bootstrapping algorithm developed within the Insemtives project [31][35][36].

The above mentioned semantic annotations and additional information gathered about a user is then used in order to offer two groups of functionalities: personalized recommendations (push) and search (pull).

Within the system twofold personalization has been applied. The first one is solely based on the information known on a single user (so called atomic personalization [15]), e.g., geo-location, outcomes from analysis of preferences based on the previous reviews. Based on the current location of a user or his interests (user profile), the recommendations of different objects to visit are to be provided. While providing the context-based recommendation, this context influences the results. However, as atomic personalization may sometimes lead to over-specialization [15], also other information is considered e.g., users who similarly rated the given object, outcomes of analysis of the friends network. This is called collaborative personalization and it can also help overcome the problem of a cold start, when little or no information is known about an individual. In addition, semantic based clustering of users is performed, where the reasoning is to be applied in order to compute the distance between different tags and users. The information on which aspects a user usually points to is also used by a system in order to perform user clustering and conduct personalized search. This is to be used while providing suggestions of places to visit and ranking search results.

As semantic annotations allow overcoming problems derived from the ambiguous nature of the natural language and from the specificity gap between annotations and queries, the system may ensure a higher precision to its users (e.g., [35]). In order to take advantage of the semantic annotations, a specific search interface is to ensure the correspondence of the user query with the underlying semantic annotation model. Therefore, the bootstrapping algorithm is also used while formulating a query by a user. In addition, the interface allows a user to formulate a query using building blocks, thus, giving him a full control on the extent of personalization and constraints used in the query. In this way, the undesired limitation of the world is avoided (e.g., localised search is desirable when a user is searching for restaurants locally, but is not as desirable when one just wants to find the best restaurant worldwide). Once a user clicks the submit button, the formulated query is resolved, and the SPARQL query is created and executed on the Insemtives platform [37] that retrieves the recommendation results. Resolving spatial-queries, recommended by my friends, or people of similar interests or having some specific aspects annotated is also supported by the Taste it! Try it! application. The user is also able to influence the ranking of search results by ranking different search dimensions.

To summarize, the application exhibits the following features: ontology-based structuring of text reviews using additional domain knowledge and taking advantage of the LOD data sets, multi-layer semantic-based user clustering and context-aware personalization of search results.

3.3 MOTIVATION LAYER APPLIED

The social aspect of user gratification is expected to solve one of the motivational problems that social software based on user contribution is facing. This is the problem of decoupling users' roles of a contributor and beneficiary of the system. A written and submitted review of a given restaurant can, of course, be valuable for its author in the future. However, he is most likely to benefit more from

contribution of other users, not from his/her own. The time of investment is then often much different from the time of benefit, this is known as the "curse of prepayment" [29].

Within the INSEMTIVES project, the issue of motivating users to contribute to the semantic content creation was investigated and relevant guidelines and models have been provided [30].

As it was already mentioned, the goal of the Taste it! Try it! is also to motivate users to create semantic annotations of restaurants. The annotations assigned to a restaurant by the application, are derived from collective decisions of reviewers. Thus, it is required to motivate users to produce reviews in considerable amounts and with a substantial level of details. The incentives models and methods within the application may be summarised according to two dimensions: usability design related incentives, such as user-friendliness and easiness of creation of semantic annotations; and sociability design manifested through the integration of the application with the Facebook portal and usage of badges and points to award users for the activities.

The first mentioned dimension of incentives relates to the design of the application and its interface. It covers such usability design aspects as controllability, self-descriptiveness, error-tolerance, expectation conformity, suitability for task and individualization. While developing the application, our main motivation was to hide the complexity of semantics being the backbone of the application. The semantic annotations that are created are template based annotations, thus, the entire process of creating annotations is more user friendly and resembles typical interaction with the Web 2.0 application. Even the creation of the semantic annotations based on the free text concepts is more user friendly thanks to the application of the bootstrapping algorithm, already mentioned.

The second groups of mechanisms include the sociability design aspects, that manifest themselves by awarding badges to users being the most active or reaching certain thresholds e.g., for each review submitted, users are awarded with points. In turn, badges show the status of a user, his/her hobby as well as current achievements. The gratification rules define when a user is eligible to get a certain badge. Both badges and points are displayed in the profile and on the wall of the user on the Facebook portal. It allows taking advantage of the following motivation levers: reputation, competition, conformity to a group, usefulness, altruism, reciprocity and self-esteem.

3.4 Architecture and the Insemtives platform

To fulfil goals defined for the application, five major components of the application were distinguished, namely: server, Android client, Facebook client, Facebook and the Insemtives platform. The Android client provides a user with a mobile front-end to manage reviews. The server component performs the semantic annotation process and publishes the prepared LOD using the Insemtives platform. The server also provides an interface for the Facebook client that enables retrieval of information on the user interactions with the application, as well as on restaurants and reviews.

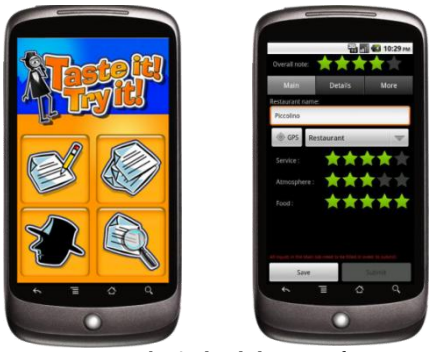


Figure 1. The Android client interface

The server also updates information on statistics, granted badges and uses the Facebook Graph API to post information on the Facebook wall of the user e.g., about a new review or a new badge granted to the user. The Facebook client is another front-end to the application and is embedded in the Facebook canvas. It uses the Facebook JavaScript API to retrieve basic information about the user including Facebook user ID, user name, friends, location, locale, etc. This data may be used in the Facebook side calls. The Insemtives platform enables publishing data in the LOD cloud, offers SPARQL support while accessing the data from the cloud, as well as provides the bootstrapping component.

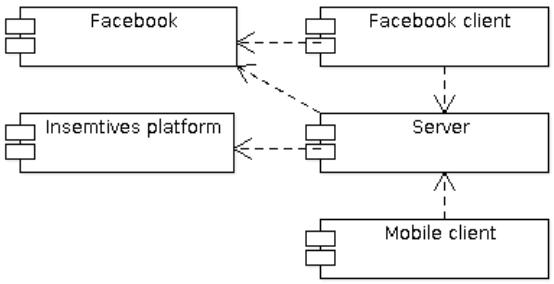


Figure 2. System schema

4. Conclusions and future works

The Taste It! Try It! application presented in this paper is to support users in creating semantically annotated reviews. This goal is achieved by providing an application similar to what users already use and applying incentives mechanisms to motivate them. These social incentives mechanisms taking advantage of the Web 2.0 ideas are to guarantee the appropriate quantity and quality of the created semantic annotations of objects. This will in turn allow offering personalised and more accurate search possibilities leading to creation of a valuable recommendation system, thus, constituting additional incentive for users to use the application. The Taste It! Try It! Application offers the added value towards the existing recommendation systems especially in the area of personalization of search results and contextual semantic search. Worth mentioning is also the integration with the

 ISBN: 978-1-61208-181-6

LOD cloud. We believe that features of the Taste It! Try It! application provide a reasonable compromise between functionality, usability, simplicity and attractiveness from the user point of view. However, only an evaluation of the proposed solution being a part of our future work will show, whether the application constitutes a good compromise between the power of semantic annotations and difficulty of creating and maintaining them.

ACKNOWLEDGMENT

This work has been partially funded by the FP7 project INSEMTIVES, EU Objective 4.3 (grant no.FP7-231181).

REFERENCES

[1] Adomavicius, G., "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions", IEEE Transactions on Knowledge and Data Engineering, vol 17, no. 6, 2005.

[2] Yang, W., Wang, Z., and You, M. "An improved collaborative filtering method for recommendations' generation", in Proc. of IEEE International Conference on Systems, Man and Cybernetics, 2004.

[3] Carrillo de Albornoz, J,, Plaza, L., Gervás, P., and Díaz, A.," A joint model of feature mining and sentiment analysis for product review rating". In Proceedings of the 33rd European conference on Advances in information retrieval (ECIR'11), Paul Clough, Colum Foley, Cathal Gurrin, Hyowon Lee, and Gareth J. F. Jones (Eds.). Springer-Verlag, Berlin, Heidelberg, 2011, pp. 55-66.

[4] Peis, E., J. M. MMorales del Castillo, J.M., and Delgado-Lpez, J. A., . Semantic recommender systems. analysis of the state of the topic,. Hipertext.net, no. 6, 2008.

[5] Gartner's Report, Emerging Trends and Technologies Roadshow for 2008 to 2012, 2007, last access date: 13.09.2011.

[6] Berners-Lee, T., Hendler, J., and Lassila, O., (2001). The semantic web. "Scientic American", no. 284(5) 2001, pp. 34-43.

[7] Uschold, M. and Gruninger, M., Ontologies: principles, methods, and applications, "Knowledge Engineering Review", no. 11(2) 1996, pp. 93-155.

[8] Shadbolt, N., Berners-Lee, T., and Hall, W., The semantic web revisited, "IEEE Intelligent Systems", no.21. 2006, pp. 96-101.

[9] Siorpaes, K. and Simperl, E., Human intelligence in the process of semantic content creation, "World Wide Web Journal", no. 13(1), 2010, pp. 33-59.

[10] http://insemtives.eu last visited: 13.09.2011

[11] Burstein, F "Systems Development in Information Systems Research". In: Williamson, K(ed.), Research Methods for Students, Academics and Professionals: Information Management and Systems, 2nd edition. Wagga Wagga, New South Wales, 2002

[12] Bizer, C., Heath, T., and Berners-Lee, T., Linked data - the story so far., " International Journal on Semantic Web and Information Systems", no. 5(3) 2009, pp. 1-22.

[13] http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData - last visited 13.09.2011

[14] Resource Description Framework - http://www.w3.org/RDF/, last visited: 13.09.2011

[15] Garcia-Molina, H., Koutrika, G., Parameswaran, A., "Information Seeking: Convergence of Search, Recommendations and Advertising, Communications of the ACM, 2009

[16] Pazzani M. and Billsus D.. "Learning and revising user profiles: The identification of interesting web sites". Machine Learning, 27:313-331, 1997.

[17] O'Reilly, T. and Battelle, J., "Web squared: Web 2.0 five years on", 2009.

[18] Breese, J.S., Heckerman, D., and Kadie, C.. "Empirical analysis of predictive algorithms for collaborative filtering". In 14th UAI Conf., 1998.

[19] Hofmann, T.. "Latent semantic models for collaborative filtering". ACM TOIS, 22(1), 2004.

[20] Micarelli, A., Gasparetti, F., Sciarrone, F., and Gauch, S. "The Adaptive Web", volume 4321 of LNCS, chapter Personalized Search on the World Wide Web. 2007.

[21] Bizer, C., Cyganiak, R., and Gauß, T.: The RDF Book Mashup: From Web APIs to a Web of Data. Proceedings of the 3rd Workshop on Scripting for the Semantic Web , 2007

[22] Heath, T. and Motta, E. "Revyu: Linking reviews and ratings into the Web of Data". Journal of Web Semantics, 6(4):266-273, 2008.

[23] Aciar, Silvana, Zhang, Debbie, Simoff, Simeon, and Debenham, John. Recommender System Based on Consumer Product Reviews. In Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI '06). IEEE Computer Society, Washington, DC, USA, 719-723.

[24] Sugiki, Kenji and Matsubara, Shigeki."Product retrieval based on semantic similarity of consumer reviews to natural language query". Int. J. Knowl. Web Intell. 1, 3/4 (July 2010), 209-226.

[25] Benkler, Y. The Wealth of Networks: How Social Production Transforms Markets and Freedom. Yale University Press, New Haven, CT, 2007.

[26] Malone, T. W., Laubacher, R. Dellarocas, C., Harnessing Crowds: Mapping the Genome of Collective Intelligence. MIT, City, 2009.

[27] von Ahn, L. Games with a purpose. Computer, 39, 6, 2006, 92-94

[28] Nov, O., Naaman, M. and Ye, C. What drives content tagging: the case of photos on Flickr. In Proceedings of the Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems (Florence, Italy, 2008). ACM

[29] Zacharias, V. and Braun, S. (2008). Tackling the curse of prepayment collaborative knowledge formalization beyond lightweight. In 1st Workshop on Incentives for the Semantic Web , 7th International Semantic Web Conference ISWC2008, October 27th, 2008, Karlsruhe, Germany, CEUR Workshop Proceedings

[30] Cuel, R., Tokarchuk, O., and Zamarian, M.: Mechanism Design for Designing Annotation Tools. Proceedings of Sixth International Conference on Internet and Web Applications and Services, St. Maarten, The Netherlands Antilles, March 20-25, 2011.

[31] D4.3.2Bootstrapping tools for image files (final version), Insemtives Project FP7-ICT-2007-3, 30.03.2011; http://www.insemtives.eu/deliverables/INSEMTIVES_D4.3.2.pdf; last visited: 13.09.2011

[32] Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G., and Reyna, J. "Building a Sentiment Summarizer for Local Service Reviews", r, WWW Workshop on NLP Challenges in the Information Explosion Era (NLPIX), 2008.

[33] Beineke, P., Hastie, T., Manning, C., and Vaithyanathan S.. "An exploration of sentiment summarization". In Proceedings of National Conference on Artificial Intelligence (AAAI), 2004

[34] Hu M. and Liu. B. "Mining opinion features in customer reviews". Proceedings of National Conference on Artificial Intelligence (AAAI), 2004

[35] Andrews, P., Pane, J., and Zaihrayeu, I., "Semantic Disambiguation in Folksonomy: a Case Study". to appear in Advanced Language Technologies for Digital Libraries, Springer's Lecture Notes on Computer Science, Hot Topic subline, 2011

[36] Andrews, P., Kanshin, S., Pane, J., and Zaihrayeu, I.:"Semantic Annotation of Images on Flickr". Demo Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011), Springer LNSC, Heraklion, Greece, May 29th - June 2nd, 2011

[37] D3.2.2 Deliverable - Semantic Content Management platform (final version), Insemtives Project FP7-ICT-2007-3, 30.03.2011; http://www.insemtives.eu/deliverables/INSEMTIVES_D3.2.2.pdf; last visited: 13.09.2011.

 ISBN: 978-1-61208-181-6

Conceptual Modeling in Wikis: a Reference Architecture and a Tool

Chiara Ghidini, Marco Rospocher, 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 organisational 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 describe MoKi, a conceptual modeling wiki for ontologies and business processes fully compliant with the presented reference architecture.**

***Keywords*-Conceptual Modeling, Collaborative Modeling, Semantic Wikis, Ontology Modeling, Process Modeling**

I. INTRODUCTION

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. While traditional wikis allow to enter unstructured text and multimedia content directed to other human users, and not in a format apt to be understood by computers, recent projects, such as DBpedia [1], YAGO [2], and Semantic Media Wiki (SMW) [3] have empowered traditional wikis with the capability of publishing their 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 (online) 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+ [4], MoKi [5], and Ontowiki [6]). and has suggested the usage of the wiki philosophy in tools which are not directly built on top of wikis (e.g., Senso Comune [7], Freebase [8], and PoolParty [9]).

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. 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, 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 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 scenarios composed of different models. *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 [10], 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.*

The contribution of this paper is twofold. 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

 ISBN: 978-1-61208-181-6

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 both by domain experts and knowledge engineers. Second, we illustrate an implementation of this architecture in MoKi, a wiki for modeling ontologies and business processes. This implementation aims at showing the feasibility of the architecture by means of a practical realization.

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 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.

The paper is structured as follows: we start from an analysis of conceptual modeling languages (§II) and we proceed by defining an architecture which satisfies the needs of generality and collaboration (§III and §IV). We then provide a description of MoKi (§V) and we conclude with a comparison between the proposed architecture and state of the art tools for wiki-based conceptual modeling (§VI).

II. CONCEPTUAL MODELING

Conceptual modeling (aka semantic modeling) has been researched into and used in several areas of Computer Science and Engineering often with different usages, characterizations, and terminologies. According to [11] and [12], 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, [12] 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. They are used to describe different types of concepts, such as Entities, Activities, Agents, Goals, and so on, depending on the CLM used; and
2) **Organisational mechanisms**: these are primitive mechanisms for structuring the model along different dimensions. Examples of organisational mechanisms (also called *abstraction mechanisms* in [12]) are generalization (often referred to as *isA*), aggregation (*partOf*), classification (*instanceOf*), contextualisation / modularization, and so on.

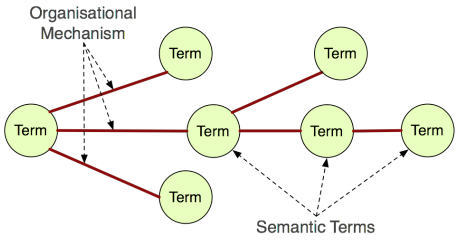


Figure 1. Conceptual Modeling Languages.

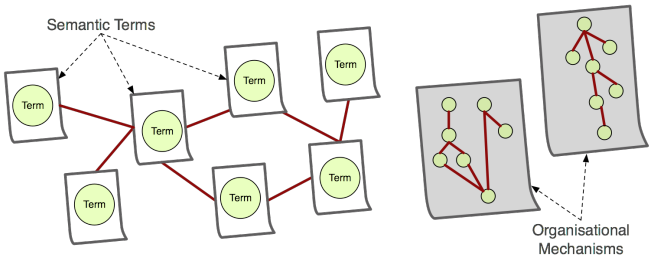


Figure 2. Representing a conceptual model in a wiki.

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 extremely informal semantics (in some cases even none) is usually associated to the diagrams. On the contrary, if the models need to be as less ambiguous as possible, 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 there are "semi-formal" CMLs: an example is the Business Process Modeling Notation [13], 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 organisational mechanisms. In this section we introduce the notion of Conceptual Modeling Wiki (CMW) which uses wiki pages to represent these building blocks.

A pictorial representation of a CMW is given in Figure 2. A CMW 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 organisational mechanism. For instance, if

 ISBN: 978-1-61208-181-6

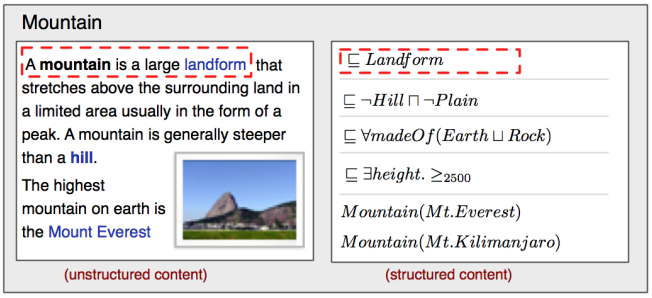


Figure 3. Wiki page for semantic terms.

we consider a CML having as semantic terms concepts, instances, and roles, and two organisational mechanisms such as generalisation 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 generalisation and the aggregation/decomposition dimensions respectively.

A. Building Wiki Pages for Terms

The idea of associating a wiki page to each semantic term is adopted by most of the state of the art wiki-based tools used to represent and manage knowledge (see §VI). 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 at http://en.wikipedia.org/wiki/Mountain, 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 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, 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.

The advantage of storing the unstructured and structured descriptions within 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 organisation of a page in an unstructured and structured part is a second important step in defining the architecture of a CMW, but may not be enough in the case of complex CMLs, such as the ones based on logical formalisms (e.g., OWL [14]) or very complex notations (e.g., BPMN [13]). In this case the structured part of the page will contain very precise (often logic based) description of a term, preventing domain experts from accessing the domain knowledge encoded in the conceptual model.

 ISBN: 978-1-61208-181-6

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.

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 (structured) knowledge. Examples are: different access modes which represent the OWL structured content using different syntax, controlled natural languages, or graphical representations. Analogously 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 or complex 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.

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 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 is a collaborative, MediaWiki-based [15], 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 any instantiation of the 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 an implementation of MoKi (see also [5]), fully compliant with the architecture illustrated in §III–IV. A running installation of MoKi can be tested on-line at https://moki.fbk.eu/moki/tryitout2.0.

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* (we use this term 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 §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.

 ISBN: 978-1-61208-181-6

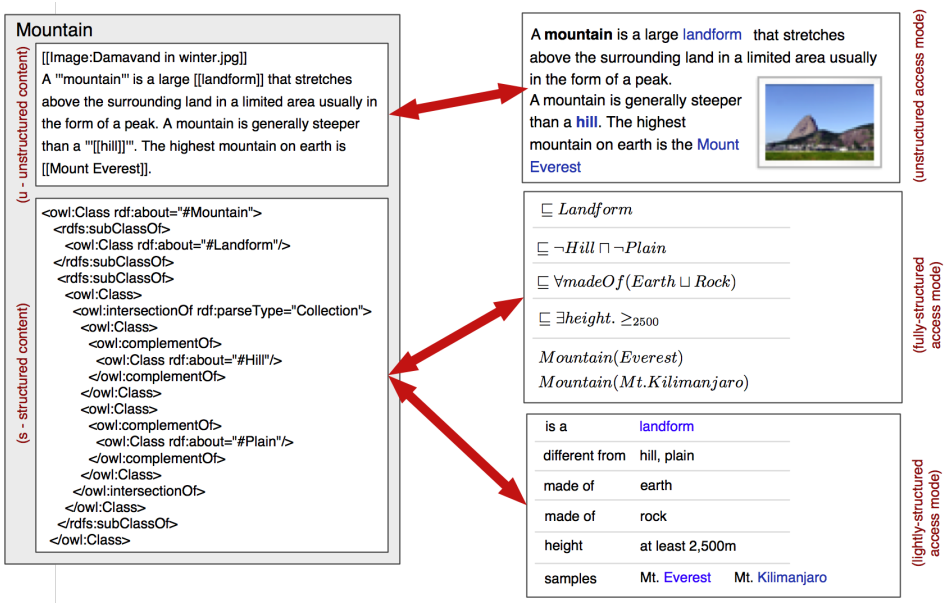


Figure 4. Multi-mode access to a wiki page for semantic terms.

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 serialisation of a set of OWL statements formalising the term, while, similarly, the structured part of a page describing a BPMN process contains an XML serialisation of the JSON object representing the process diagram.

B. *Multi-mode access in* MoKi

Users can access the ontological and procedural knowledge contained in MoKi using the three different access modes described in §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.

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, an intuitive latex-style format for writing ontologies using a text-editor, format which can be automatically translated into (an RDF/XML serialisation of) OWL. The latex2owl syntax was chosen because of its resemblance to the DL syntax, but the tool can be adapted to handle fully-structured access mode based on other OWL syntaxes such as the Manchester OWL syntax. The user can easily edit the list of axioms in a form based interface, as the one shown in the top part of Figure 5. 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 the top part of Figure 6. In particular we have tightly integrated in MoKi the Oryx editor [16], a full-fledged editor that allows to create processes according to several modeling languages, including BPMN.

The lightly-structured access mode: As described in §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 the bottom part of Figure 5. 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 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

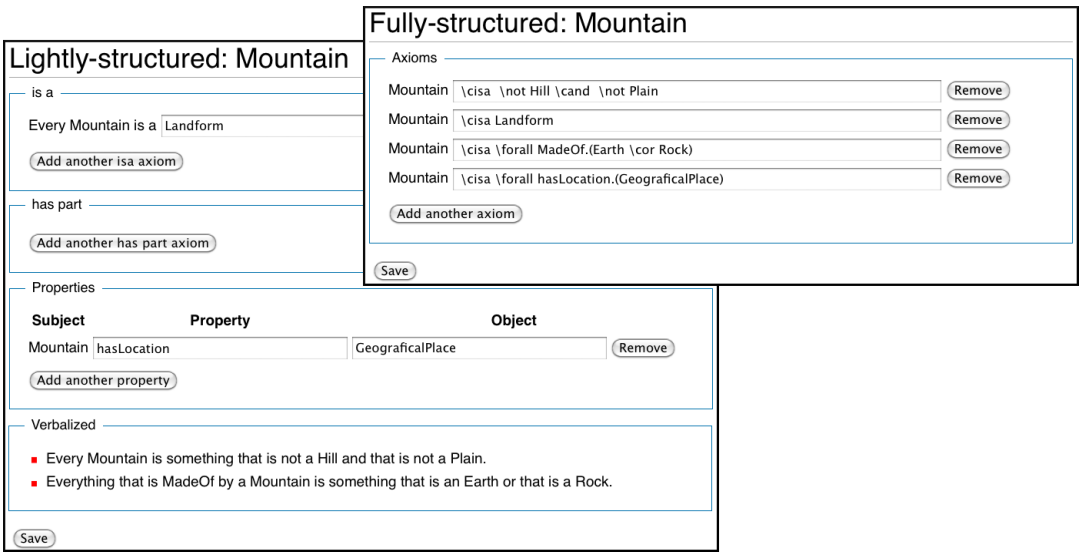


Figure 5. Fully-structured access mode and lightly-structured access mode to the page of concept *Mountain*.

\forall *property*.(*object*)". 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 [17]. 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 the Oryx editor (see the bottom part of Figure 6) which shows only the basic workflow of the activity, the main elements of the process such as start and end events and the (sub-)processes it can contain, hiding the details and complexity typical of BPMN diagrams.

C. *Organisational mechanisms in* MoKi

Organisational 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 organisational mechanism rely also on graphical forms of representation, which include graphical browsing and editing facilities. For ontological knowledge the organisational mechanism pages allow to explore and edit the generalisation and part/subparts decomposition hierarchies of ontology concepts, as well as the classification of the ontology individuals. For procedural knowledge, the current organisational mechanism pages provide an overview of the activity/sub-activity decomposition, and 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.

VI. 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.

Focusing on tools, wiki systems and semantic wikis have been mainly applied to support collaborative creation and sharing of ontological knowledge. *AceWiki* [18] was developed in the context of logic verbalisation, 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+* [4], 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* [19] supports

 ISBN: 978-1-61208-181-6

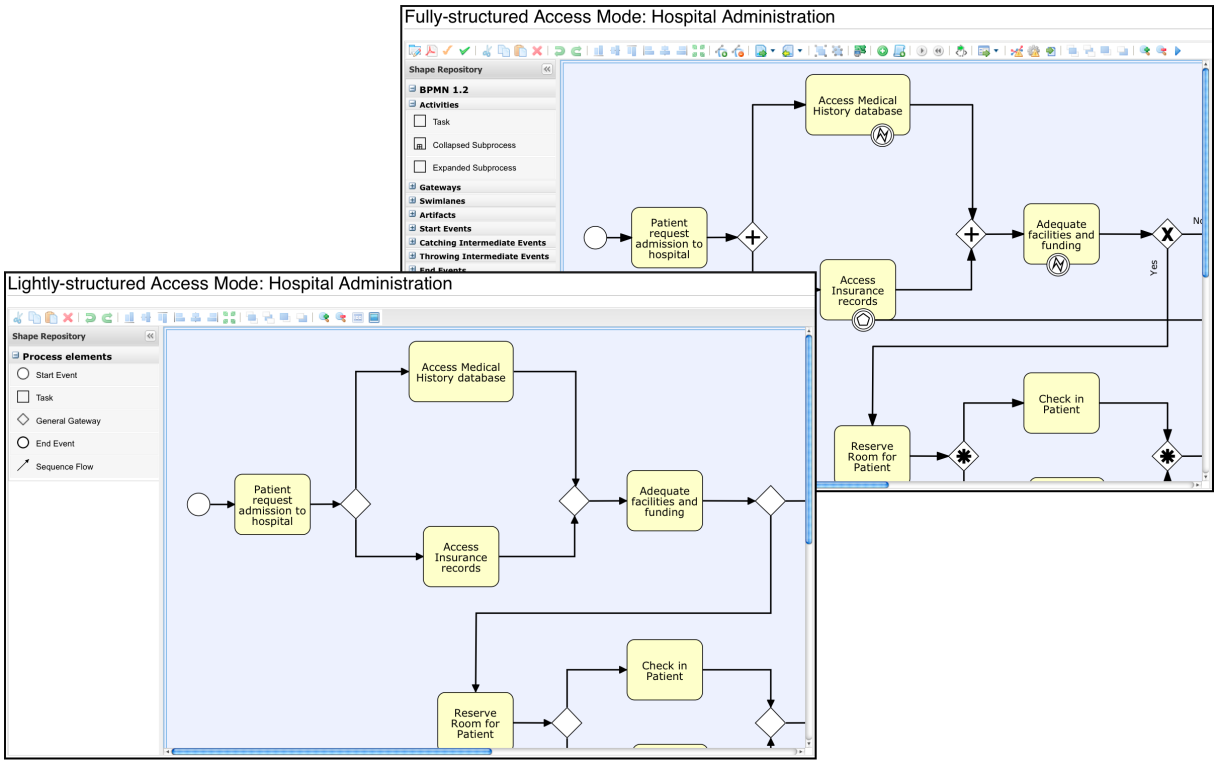


Figure 6. Fully-structured access mode and lightly-structured access mode of a process.

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* [6] 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 workflows using Semantic MediaWiki is implemented in the *Semantic Result Formats* extension [20].

We have compared the tools mentioned above, together with the current and previous versions of MoKi (a preliminary version of the tool was presented in [21]), against the distinctive characteristics of our reference architecture. The results are displayed in Table I, where the columns 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 organisational 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*).

As we can see from the table, 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.

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

	1 page/ 1 term	overview	unstruct./ struct.	multi-mode	multiple CMLs
AceWiki	X				
SMW+	X	X	X		
IkeWiki	X		X		
OntoWiki	X	X	X		
Sem. Res. Form.	X	X	X		
MoKi v.1	X	X	X		X
MoKi v.2	X	X	X	X	X

VII. CONCLUDING REMARKS

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 an implementation of MoKi fully compliant with the proposed architectural model.

A customized version of MoKi described in this paper is being successfully used by domain experts in five Italian

 ISBN: 978-1-61208-181-6

regions within the ProDE National project to develop models of documental flows in five different sectors of the Public Administration (PA). The models are composed of a process model, describing the flow of activities carried out in a PA sector, and of a domain ontology, describing the documents generated / used by the different activities, and the roles performing the different activities. A qualitative evaluation on the entire modeling process has been performed [22], and the results shows that the domain experts perceived the tool as more than *easy to use*, and *useful* for the collaborative modeling of documents and processes. A customized version of the tool, called *CliP*-MoKi, has also been applied to model clinical protocols encoded in the ASBRU language [23].

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. 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.

REFERENCES

[1] 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 WWW*, vol. 7, no. 3, pp. 154–165, 2009.

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

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

[4] 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.

[5] C. Ghidini, M. Rospocher, and L. Serafini, "MoKi: A Wiki-Based Conceptual Modeling Tool," in *Proc. of ISWC 2010, Posters and Demonstrations Track*, Shanghai, China, 2010.

[6] 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.

[7] A. Oltramari and G. Vetere, "Lexicon and ontology interplay in senso comune," in *Proceedings of OntoLex 2008*, Marrakech (Morocco), 2008.

[8] 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.

[9] 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.

[10] V. Dimitrova, R. Denaux, G. Hart, C. Dolbear, I. Holt, and A. G. Cohn, "Involving domain experts in authoring owl ontologies," in *Proc. of ISWC 2008*, ser. LNCS, vol. 5318. Springer Berlin / Heidelberg, 2008, pp. 1–16.

[11] 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.

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

[13] OMG, "Business process modeling notation, v1.1," www.omg.org/spec/BPMN/1.1/PDF, January 2008.

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

[15] Wikimedia Foundation, "Mediawiki," http://www.mediawiki.org, last Accessed on 6 Nov 2011.

[16] The Oryx Project, "The oryx editor," http://bpt.hpi.uni-potsdam.de/Oryx/, last Accessed on 6 Nov 2011.

[17] K. Kaljurand and N. E. Fuchs, "Verbalizing owl in attempto controlled english," in *Proceedings of Third International Workshop on OWL: Experiences and Directions, Innsbruck, Austria (6th–7th June 2007)*, vol. 258, 2007.

[18] T. Kuhn, "AceWiki: A Natural and Expressive Semantic Wiki," in *Proceedings of Semantic Web User Interaction at CHI 2008: Exploring HCI Challenges*, 2008.

[19] S. Schaffert, "Ikewiki: A semantic wiki for collaborative knowledge management," in *1st Int. Ws. on Semantic Technologies in Collaborative Applications (STICA'06)*.

[20] 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.*

[21] M. Rospocher, C. Ghidini, V. Pammer, L. Serafini, and S. Lindstaedt, "Moki: the modelling wiki," in *SemWiki2009 - Fourth Workshop on Semantic Wikis*. CEUR-WS.org, 2009, pp. 113–127.

[22] 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, 2011, vol. 7032, pp. 17–32.

[23] 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.

 ISBN: 978-1-61208-181-6

Reorganization of KM-Oriented Medium Voltage Power System Planning Process

Ricardo H. Guembarovski
Post-Graduate Program in Knowledge Engineering and Management
Universidade Federal de Santa Catarina (UFSC)
Florianópolis-SC, Brazil
E-mail: ricardohg@celec.com.br

Jose Leomar Todesco
Post-Graduate Program in Knowledge Engineering and Management
Universidade Federal de Santa Catarina (UFSC)
Florianópolis-SC, Brazil
E-mail: tite@egc.ufsc.br

Murialdo Loch
Post-Graduate Program in Knowledge Engineering and Management
Universidade Federal de Santa Catarina (UFSC)
Florianópolis-SC, Brazil
E-mail: murialdo@egc.ufsc.br

Jeferson de Souza
Post-Graduate Program in Knowledge Engineering and Management
Universidade Federal de Santa Catarina (UFSC)
Florianópolis-SC, Brazil
E-mail: jefersons@celesc.com.br

***Abstract*—**The planning of medium voltage power systems has been so far carried out in a limited way. Apart from the purely technical aspects, other issues as relevant as the task of planning must be identified and defined. The methods currently available are based solely on mathematical principles and computational techniques applied to the schematic representation of power systems. The organization of the planning process proposed here stems from a cognitive approach that associates intrinsic knowledge with planning activities and other relevant aspects. A description of the organization of the planning process is preliminary presented so that the other aspects inherent to this activity are identified. From a systemic perspective, the re-organization of the planning process is proposed taking its efficacy into account. Finally, the reorganization of the planning process is evaluated so that its properties are identified along with its efficiency. Undoubtedly, the planning of medium voltage power systems requires improvement. The optimization of this process transcends classical and purely technical problems with power systems, which leads us to propose a reorganization of the planning process, focusing on knowledge management (KM) as the main paradigm of investigation.

Keywords-Process Management; Planning of Power Systems; Knowledge Management

I. INTRODUCTION

This paper discusses the planning of electric power distribution systems, more specifically the planning of a medium voltage distribution system. The re-organization of this planning process based on knowledge management is the main purpose of this study.

The Brazilian national electric power industry has been considered one of the best in the world in terms of reliability and operational costs [1]. With some rare exceptions, the electrical power distribution system has always presented supply quality levels compatible with those demanded from the consumer market.

In general, the interconnected system receives the energy generated by hydroelectric, eolic and thermal plants. The whole process occurs in accordance with rules established by various regulating agencies. In this context, Agência Nacional de Energia Elétrica – ANEEL [National Agency for Electrical Energy] has as a mission to provide favorable conditions so that the electrical energy market develops in a sustained and balanced way among its agents and for the benefit of society [2].

The electrical power system consists of three interconnected components, each one with very distinct features: generation, transmission and distribution. The distribution systems, so called due to the fact they operate on voltage equal to or lower than 138 kV, distribute energy to all classes of consuming clients.

In accordance with PRODIST [3] the distribution system can be classified as: High Voltage Distribution System (HVDS) with nominal voltage between 138 kV and 69 kV; Medium Voltage Distribution System (MVDS) with nominal voltage between 13.8 kV and 34.5 kV; and Low Voltage Distribution System (LVDS) with nominal voltage between 440 Volts and 110 Volts. The distribution systems are interconnected by means of distribution substations (DSs) which aim to transfer vast amounts of power at more adequate voltage levels to distribution in accordance with urban specificities. Fig. 1 displays an electric power distribution system.

The MVDS, a primary network segment in Fig. 1, aims to supply electric power from the DSs to low and medium voltage clients, which include large companies, industries, commercial clients and residences located in the rural and urban areas. This preliminary study focuses exclusively on the planning of a MVDS.

It is argued here that the planning activities carried out by electric utilities need improvement, since there are difficulties in meeting regulatory goals concerning supply quality problems and in justifying investments. Ultimately, a planning activity requires approaches suitable for the scientific paradigms that support the management processes of modern organizations.

 ISBN: 978-1-61208-181-6

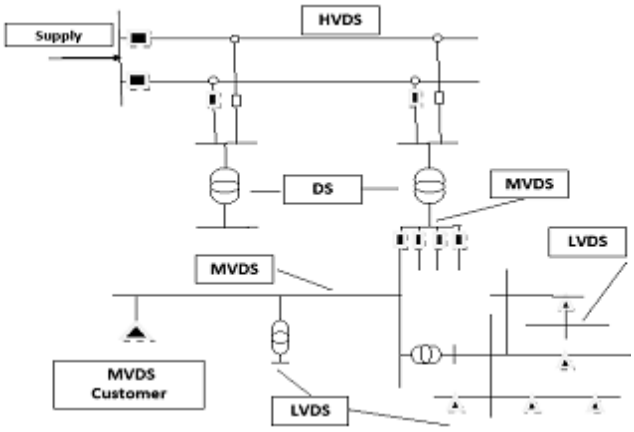


Figure 1. Representation of an electric power distribution system.

In this context, it is important to highlight that the organizations that make use of old-fashioned paradigms show organizational deficiencies in which the main reasons can be [4]:

- Lack of focus on the client;
- Lack of clear, well-defined and disseminated objectives and concepts;
- Processes and activities are not registered or optimized;
- Professionals who do not know the organization's role and do not participate of processes, actions and solutions to problems;
- Lack of ways to constantly measure and evaluate the processes.

Although there are good prospects for the quality of electricity supply in relation to a medium voltage distribution system, there is much to be done in terms of improving the planning process:

a) operationally, long and successive energy supply interruptions and problems of voltage regulation occur in MVDSs and LVDSs, resulting in damages and regulatory fees. Additionally, these problems detract the company's image in a significant way.

b) tactically, organizational problems associated with indefinite processes often occur. There is some difficulty in integrating and standardizing procedures and Technologies.

c) strategically, planning experts are questioned about their reasons for technical alternatives and requests for resources, which are not convincingly justified. There is much uncertainty and generally decision-makers, who act strategically, do not obtain the necessary information to make effective decisions due to lack of appropriate computational tools and processes.

From a historical perspective, investments have not followed the demands for market expansion. In addition, experts face difficulties in articulating good investment justifications; the regulatory rigor is strong; financial resources are short; consumer market is demanding; ultimately, decision-makers need to know the organizational processes, that is, they need knowledge. It is believed that these issues would be attended to if a reorganization of the MVDS planning process was effectively implemented and supported by appropriate knowledge management.

In line with [23], R$ 75 million was spent nationwide on fines resulting from problems of power quality supply between 2008 and 2009. Furthermore, it is highlighted that some electric utility companies going in the opposite direction of the regulatory framework signal an investment reduction, which implies a risk to distributors. ANEEL shows that investments must be justified and when approved they must be carried out in their integrity under penalty of the utility companies being discounted in the next tariff cycles [23].

In order to acquire a real picture of this process, in 2009, CELESC Distribution S.A., a state electric power distribution company in Santa Catarina state, provided electric power to approximately 2,256,178 consumers and invested more than R$ 66 million in LVDS and MVDS distribution networks [21].

In order to stay competitive in the market, the electricity distribution companies need systematically develop an investment plan consistent with reality. Technical data, information about electric systems and the market, ultimately, knowledge of this process is indispensable. In this sense, [18] affirm that within information economy, pivotal knowledge-based competencies represent the organization's greatest asset. Knowing how to do things effectively is the greatest differentiator of success.

Another fundamental aspect when organizational issues are discussed has to do with the need to evaluate all the component parts of a process from various points of view. According to [24], systems thinking enables professionals to understand that a process consists of parts dependent on the whole and that these parts should not be analyzed in isolation. Systems thinking leads us to always evaluate the whole, considering the other disciplines of which it is comprised.

In Sections 2 and 3, the current process of planning a power system is presented along with a brief introduction to knowledge management connected with the systems view, which is an intrinsic foundation to the study being developed. In Sections 4 and 5, the reorganization of the planning process is proposed with a view to knowledge management. Finally, conclusions and recommendations are offered.

II. THE CURRENT MVDS PLANNING PROCESS

Organizational efficiency has been an object of research due to its relevance to the economic context. Production with the highest possible quality and lower costs is the goal of any company that intends to stay competitive in the market.

According to [13], as large companies perceive that one of the greatest competitive advantages of organizations is the production of intangible assets, knowledge of business processes becomes increasingly critical. In [12] view, these processes represent to modern organizations the essence to enable their existence, especially when the analysis of an organization emphasizes the evaluation of its processes and not only the results of these processes.

 ISBN: 978-1-61208-181-6

Still according to [13], much of the literature defines a process as a set of sequential tasks that receive input, process value-adding functions, and provide customers with a product or service. Therefore, an organization has within itself a set of processes that constitute it, and some processes can be more efficient and effective than others.

With a vast bibliography, the planning of distribution systems counts on a set of methods and computational techniques that have been developed for decades. Module 2, which refers to the PRODIST, establishes the guidelines for the expansion planning of the distribution system. Either for a regulatory or financial reason, the level of demand increasingly imposes that electric utility companies continue to improve their results.

Along with [11], the expansion planning of distribution systems consists of proposing, analyzing and selecting expansion alternatives to meet the increasing demand, respecting guidelines, restrictions and the criteria for the quality of electric power supply. The lower cost alternatives, which meet the established criteria, are selected and integrated into a work plan for the study period here established.

Generally speaking, the planning of distribution systems can be understood as follows [10]: Be it a distribution system meeting the demands of an electric power market comprised of consumers. The demand associated with electric power consumption is dynamic and varies in space and time. The increase of demand requires the expansion of the distribution system that can be translated in general lines by the following actions: build substations, increase the capacity of transformation, build new lines and/or change lines (i.e. reconducting). Develop a work plan taking into account network operation costs, and at the same time meeting a set of regulatory, economic and operational restrictions constitutes a planning problem.

Various studies have dedicated to the analysis of distribution network planning ([14][15][16][20]). Most of these studies, however, have an exclusive focus on the development of mathematical and algorithmic models based on schematic representations of a power system, which by considering restrictions and technical criteria enable experts to propose solutions to the identified problems. The search of an "optimal solution" is always a goal to be reached; however, the kinds of work to be carried out are selected most of the time by considering planning experts' knowledge and intuitions.

The MVDS planning is carried out considering five years of annual periodicity. The following study year receives special attention from experts. Additionally, from a purely technical perspective, experts consider the demand forecast, criteria and studies on planning, in accordance with the procedures set out in PRODIST.

More specifically, the MVDS planning has been carried out along with electric utility companies in a matrix and analytical way. Generally, the experts identify the technical problems, analyze information and after technical discussions with experts from other areas that integrate the electric power distribution system, power flow studies are carried out. According to the available budget, computational tools and mathematical methods for analysis of variables, works with the best cost-benefit ratios are prioritized.

Fig. 2 illustrates the procedures for the MVDS planning carried out along with electric power utility distribution companies.

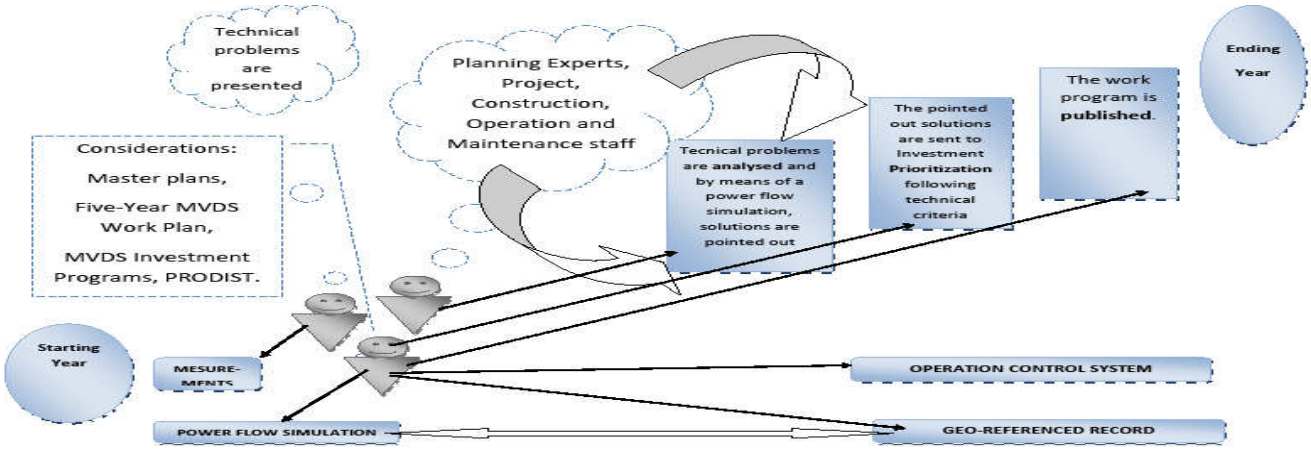


Figure 2. Representation of the current MVDS planning process.

Evaluating the current planning process (Fig. 2), some relevant issues can be identified as follows:

a) There is no distinction as to task execution. The identification of problems, development of studies and choice of alternatives are not realized in a structured way;

b) Studies are carried out, but there is no register of problems, studies and alternatives in a computational environment.

c) Conceptual indefinitions make it difficult to categorize problems, technical solutions and classify investments;

 ISBN: 978-1-61208-181-6

d) There is no structured register of studies, consequently, there is no management of the studies carried out, making it difficult to improve the process;

e) The problems are not well-defined and there are no systems to facilitate the composition of precise diagnosis as well as adequate solutions, considering the other points of view that integrate the MVDS technical problems.

Still in relation to Fig. 2, it is possible to observe that the planning procedure is carried out annually. Alternative-oriented studies are carried out, but they are not recorded in a computational environment. The computational tools are specific for the optimization and they are not integrated into a management environment.

Planning depends on other factors. The mathematical procedures currently diffused and associated with computational tools are not sufficient for the planning composition to propitiate an investment plan that is really efficient and consistent with reality.

The management of works, the supervision of the power system according to its dynamics, consensus in the selection of concepts for the identification of technical problems, structured record of proceedings, the diagnosis, and definition of problem-solution patterns; ultimately, all kinds of knowledge related to planning development are relevant for an efficient work plan to occur.

In line with [8], one of the most difficult things to be understood at present is the fact that if we do something that is good, continue to do that does not necessarily mean to get the best solution. In other words, companies cling to the paradigms that have kept them alive until the present time, safe in the knowledge that this will enable them to face new market challenges, especially competition increment and regulatory scrutiny.

III. KNOWLEDGE MANAGEMENT (KM)

In the mid-90's, knowledge management emerged as a key concept to organizations, as the basic economic resource was no longer the capital, but knowledge [9].

For Rodrigues and Helena [22], knowledge society is based on the value of intangible assets, which makes it imperative for companies to focus on knowledge management. Companies must also create ways to accumulate intangible assets, produce knowledge, transfer it, and also recognize the kind of knowledge that adds value to the company. In this scenario, organizational and entrepreneurial management must comprehend the concept of knowledge and make an effort to structure all the necessary activities to deal with the organization's intangible assets. It is, then, essential to understand how knowledge is built.

According to [19], there are two types of knowledge: tacit knowledge and explicit knowledge. Tacit and subjective knowledge kinds of knowledge are skills inherent to a person. It is a system of ideas, perception, experience that is difficult to be formalized, transferred and explained to another person. Explicit knowledge, on the other hand, is relatively easy to codify, transfer and re-utilize, as it is formalized in texts, graphs, tables, figures, drawings, schematic representations, diagrams, etc, which are easily organized in data banks and general publications either in hardcopy or electronic format.

In accordance with [25], an adequate experience of KM systematization must consider that knowledge exists in two formats: (i) people's mind and (ii) various kinds of records; therefore, information technology has a relevant part on KM, which involves its formalization, refinement and sharing. According to [19], there are four modes of knowledge conversion: (i) socialization (from tacit knowledge of an individual to another); (ii) externalization (from tacit knowledge to explicit knowledge); (iii) combination (from explicit knowledge of an individual to a group) and (iv) internalization (from explicit knowledge to tacit knowledge).

Nevertheless, implementing knowledge management requires cultural change. Apart from recognizing knowledge as an object of inestimable value, experts still need to perceive all aspects related to the vital processes of organizations. In this sense, a systemic approach stands out among others due to its peculiar ubiquity and synthesis. In view of the complexities involved, understanding the whole through a systemic approach is indispensable [6].

In this context, two postulates [6] can be highlighted:

1) Everything that exists at an abstract or concrete level is systems, components or potential components of a system; and

2) A system must be understood as a set of interconnected components that relate to each other in order to achieve a purpose.

Seen as a system, an organization can be evaluated as a system consisting of various subsystems. It can also be highlighted that systems as well as complex problems should not be analyzed in isolation as the whole is always larger than a sum of its components and presents a systemic characteristic that its components do not have [6].

Moreover, [5] explains that in any field of knowledge current problems have become very complex; solutions require interdisciplinary and systemic approaches. Like other systemic authors, [5] also warns that problems should be analyzed in isolation as specific parts of a process does not enable us to know effectively a process or complex problem. Nevertheless, if the set of components of a system and the relationship between them are known, then high levels of understanding are obtained by means of the systems [5].

Therefore, by analogy it is assumed that the process consist of components, have specific functions and systemic features. The organizational change with a view to knowledge management presupposes a systemic approach.

An interesting approach is still proposed by [7], focusing on the model and systems description. According to this proposal any system can be structured in accordance with the following attributes:

1) Composition: collection of component elements;

2) Environment: collection of items that are not part of a system, but act or suffer an action of any component;

3) Structure: collection of links between components and between these and other Items of the environment;

 ISBN: 978-1-61208-181-6

4) Mechanism: collection of processes that generate qualitative novelty, that is, they promote and obstruct transformations causing the emergency or submersion of the system or any of its properties.

Table1 shows some examples of CESM models, which can be natural, social, technical or mixed.

TABLE I. EXAMPLES OF CESM MODELS IN EXISTING SYSTEMS. SOURCE BASED ON BUNGE (2000)

System	**C** **Composition**	**E** **Environment**	**S** **Structure**	**M** **Mechanism**
Atom	Associated particles and fields comprising an atom.	Things (particles and fields) with which an atom interacts.	The fields that keep an atom together with environment items.	Processes of emission and absorption of light, combination, etc.
Company	Personnel and Management.	Market and government.	Work relationships between company members and between members and environment.	Activities that result in company products.
Solar System	Sun, planets and asteroids.	Milky Way Galaxy and other universe celestial bodies.	Gravitational forces.	Translational motion of components in orbits that enable the continuity of a system (with no dispersion or collapse) due to inertia.

According to this model, any system can be represented so that its relevant features are described. The technical properties as well as functions, combined with the description of components, structures, environments and mechanisms of the system, provide effective knowledge and enable to evaluate the capacity of the system to keep its basic properties or even (sub)emerge.

By analogy, as long as evaluation processes are understood as systems, a complete evaluation of the proposed system in accordance with the CESM model can be carried out [7]. Concomitantly to this approach, explanations related to the properties of the system associated with its mechanisms can qualify the efficiency of the proposed reorganization process.

IV. REORGANIZATION OF THE MVDS PLANNING PROCESS

As occurs in most Brazilian companies in the electricity distribution sector, CELESC Distribuição S.A. decentralized its operations to optimize the achievement of their business goals. Each of these operating units, located in regions with distinct cultural characteristics, has developed particular ways of interpreting, therefore achieving those goals, the same occurring with the planning process for the SDMT. Reflecting that unique circumstance, the company created an organizational culture that, according to [26], allowed to adjust its operations as a small business, even in the case of a large corporation. On the other hand, it made implementing corporate solutions difficult, damaging the company's organization and appropriateness to the interests of the regulator (ANEEL), in particular the question of the planning process that could no longer fulfill its role.

Consistent with [27], processes must be defined and modeled concurrently to the human tasks. The process should also consider the underlying infrastructure of the organization considering the interface with users so that interactive tasks can be defined and created.

Most importantly, the planning process is crucial to the business of an electric power distribution company, and the reorganization of the planning process-oriented knowledge management is recommended.

A KM-oriented reorganization of a planning process presupposes the restructuring of a system that has knowledge as its main paradigm. Therefore, the reorganization of a planning process starts to be understood as a system with its respective components, structure, environment and mechanisms that relate between themselves in order to achieve its purposes [7].

It is important to identify the intrinsic knowledge related to the planning system and its respective components before proposing a new MVDS planning process.

Evaluating the planning process by means of a systemic approach and identifying components individually without losing track of the system as a whole. The following dimensions are highlighted:

1) People have competencies and attributions; the ability to properly plan depends directly on the experts who need motivation, continuous training and evaluation of the results of their work;

2) The process consists of activities and tasks in accordance with norms, instructions and a timetable; in addition to the planning process itself, other processes must be structured in order of formalization, implementation, refinement and dissemination of knowledge;

 ISBN: 978-1-61208-181-6

3) The technology consists of methods and techniques that are most of the time used by means of computational tools; emphasizing that the knowledge employed by the planning process requires a transactional and knowledge systems, which are essential tools to keep track of the work and search for better investment alternatives respectively.

In an individualized way, components and their respective main attributes are highlighted in terms of the three aforementioned dimensions.

TABLE II. SYSTEMIC VIEW OF THE MVDS PLANNING SYSTEM.

Dimensions	MVDS Planning System	
	Components	Attributes
People	experts, engineers, managers and directors	motivation, tacit knowledge, explicit knowledge, attributions and established functions
Process	activities, tasks	PRODIST, planning instruction, budget, master plan, environment norms, calculation of losses, specific resolutions, technical notes
Technology	computational tools, computational agents	optimization methods, prioritization techniques, techniques for evaluation and market projection, artificial intelligence, knowledge agents, ontologies, data warehouse, transactional computer systems, knowledge systems

Table 2 shows that the planning system consists of the following components: experts, engineers, managers, directors, activities, tasks, computational tools and knowledge agents. For each of these components, their respective attributes are observed.

It is highlighted that an adequate planning presupposes the development of a work plan optimized and consonant with budget restrictions. Moreover, the planning shall consider federal, state and municipal norms and carry out studies following optimization techniques in accordance with the systemic approach, which enables to propose solutions considering all issues that influence the planning context. Economic scenarios, weather aspects, future regulatory rigor, new technologies; ultimately, all issues that direct or indirectly influence the planning environment must be systematically taken into consideration.

Moreover, with the aim of proposing a new process, Fig. 3 presents the reorganization of the KM-oriented MVDS planning process.

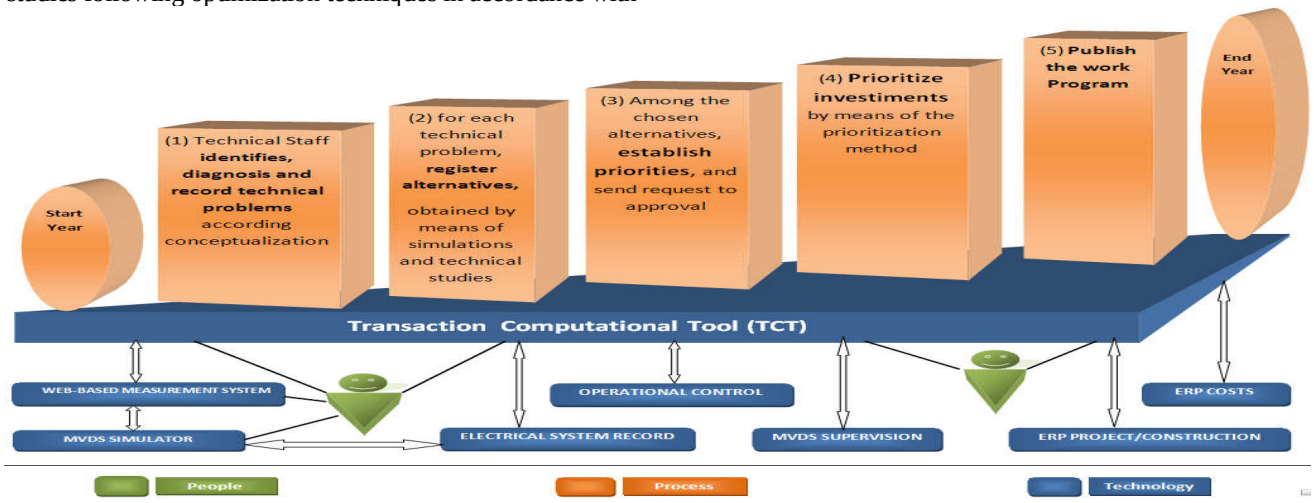


Figure 3. Reorganization of the KM-oriented MVDS planning process.

With a view to relating the dimensions identified in Table 2 to the reorganization process proposed in Fig. 3, the experts are represented in green. The relationship of experts with the process occurs by means of the TCT (Transaction Computational Tool) and demands competencies from these professional in order to carry out their activities. Technology and its components, represented in blue, are integrated with the TCT, facilitating the execution of activities, especially by adopting methods, modern techniques, dominion ontologies to provide semantic consensus, and so on. Finally, the

 ISBN: 978-1-61208-181-6

process represented in brown shows the reorganization in specific tasks that are carried out successively.

Fig. 3 shows that the MVDS planning process here proposed is structured by means of interconnected components that relate to each other in order to promote the MVDS planning. The planning task, which was previously carried out in a general way, starts to be performed in a structured way in accordance with the following sequence of tasks: identification record, problem diagnosis, studies, alternatives, prioritization, and publication of the investment plan. Moreover, norms and instructions govern and guide the execution of tasks respectively.

The TCT supports the whole process, providing experts with evaluation facilities and record of all tasks. The problems are diagnosed with the support of knowledge systems and application of methods and optimization techniques occur in an integrated way as the TCT relies on these implemented functions.

It is also important to emphasize that the TCT is critical for the reorganization of the proposed process. The interconnection of information systems that comprise the process occurs through the TCT. It is through this computational tool that the information flow occurs due to its integration with other computer systems that make up the process, like the construction management systems, operation, maintenance and ERP costs. The computational agents like knowledge system, data warehouse and artificial intelligence techniques complement the information flow in order to support the decision maker in the planning process.

Due to the reorganization of the planning system oriented towards people, processes and technology, it is noticed that the components are organized, which in turn enables the understanding of representation tasks, refinement and application of knowledge in terms of the development of an efficient work plan. It is important to highlight that the efficiency of the planning process starts to be evaluated because by means of the categorization of problem patterns, engineering actions and investments, the experts with the passing of time will be able to refine the knowledge related to the MVDS planning concomitantly with the performance of the power system.

According to [7], it is possible to verify along with the proposed system a set of mechanisms essential to the maintenance of its properties as well as its emergency condition. The CESM model is then applied to the new process in order to evaluate the reorganization of the proposed planning (see Table 3).

TABLE III. APPLICATION OF THE CESM MODEL TO EVALUATE THE REORGANIZATION OF A MVDS PLANNING.

System	C Composition	E Environment	S Structure	M Mechanism
MVDS Planning System	people computer tools, computer agents	PRODIST, MVDS, methodologies, market, building site, maintenance and operation, investors, directors, administrative council	processes, norms.	provide a work program trainings, process evaluation, motivation, formalization, representation, and refinement of the knowledge related to planning

Table 3 displays a set of mechanisms that were identified along with the proposed process reorganization. These mechanisms provide the planning system with efficiency as the investments programs start to be evaluated concomitantly with knowledge management as well as the other attributes that comprise the process. The planning system oriented towards people, process and technology facilitates the identification of the knowledge required to the execution of KM-oriented tasks.

V. CONCLUSION

The current planning process of a power system exclusively oriented to technical analysis must be re-evaluated, especially in the strategic area, which lacks compelling arguments for the development of investment plans.

Preliminarily, the current planning process was described so that its main aspects and deficiencies were identified. The reorganization of the process was proposed taking into consideration three knowledge dimensions: people, process and technology. The CESM model was used to evaluate the process reorganization that started to be understood as a system. It is important to highlight that the reorganization of the KM-oriented planning process enables the representation tasks, refinement and knowledge application to be understood. Taking into account the expertise acquired by means of the evaluation of the planning processes analyzed, experts provided with refined knowledge will effectively justify the necessary investments.

According to this proposal, the association of knowledge management with MVDS planning starts to occur in a regular way along with organizations that implement this kind of reorganization. New studies, however, will have to be consistent with the definition of knowledge systems, ontologies and motivational aspects with a view to achieving planning efficiency. It is recommended to carry out scientific research on the application of knowledge agents to the development of precise diagnosis for MVDS technical problems.

 ISBN: 978-1-61208-181-6

REFERENCES

[1] G. A. G. Santos, E. K. Barbosa, J. F. S. Silva and R. S. Abreu, Por Que as Tarifas Foram Para os Céus? Propostas para o Setor Elétrico Brasileiro. Revista do BNDES, Rio de Janeiro, V. 14, N.29, pp. 435-474, Jun. 2008.

[2] Agência Nacional de Energia Elétrica – Aneel. Obtained through the Internet: <http://www.aneel.gov.br/area.cfm?idArea=635&idPerfil=2>, (retrieved: November, 2011).

[3] ANEEL. Procedimentos de Distribuição de Energia Elétrica no Sistema Elétrico Nacional – PRODIST. Obtained through the Internet: http://www.aneel.gov.br/area.cfm?id_area=82. (retrieved: November, 2011).

[4] F. B. Batista. Governo Que Aprende: Gestão do Conhecimento em Organizações. Brasília, Jun 2004. Texto para Discussão No 1022 Governo Federal Ministério do Planejamento, Orçamento e Gestão Ministro – Guido Mantega, Secretário-Executivo – Nelson Machado, ISSN 1415-4765.

[5] L. V. Bertalanffy, 'The History and Status of General Systems Theory'. The Academy of Management Journal. Vol. 15, No. 4, General Systems Theory (Dec 1972), 407-426. Stable URL:http://links.jstor.org/sici?sici=0001-427328 197212%29 15%3A4%3C407%3ATHASOG%3E2.O.C0%3B2-4, The Academy of Management Journal is currently published by Academy of Management. (retrieved: November, 2011)

[6] M. Bunge, 'Systemism: the alternative to individualism and holism'. Journal of Socio-Economics 29 (2000) 147–157, Department of Philosophy, McGill University, 855 Sherbrooke St. W, Montreal, Quebec, Canada H3A 2T7.

[7] M. Bunge, 'How Does It Work? The Search for Explanatory Mechanisms. Philosophy of the Social Sciences'. Vol. 34 No. 2, Jun 2004, pp. 182-210, DOI: 10.1177/0048393103262550.

[8] F. Capra, (1983). O Ponto de Mutação – A Ciência, a Sociedade e a Cultura Emergente. Editora Cultrix: E.U.A .

[9] I. M. de Carvalho, 'Condições para criação de conhecimento numa organização de alta tecnologia'. In: (2006). Gestão do Conhecimento, uma estratégia empresarial, conhecer é preciso! (pp. 47-74). Brasília: J. J. Gráfica e comunicações, 2006.

[10] A. M. Cossi, "Planejamento de Redes de Distribuição de Energia Elétrica de Média Baixa Tensão". PhD thesis. Universidade Estadual Paulista – UNESP, São Paulo, Brazil 2008.

[11] D. P. Duarte, "Automação como Recurso de Planejamento de Redes de Distribuição Energia Elétrica". PhD thesis. Universidade de São Paulo – USP, São Paulo, Brazil 2008.

[12] D. A. Geuber, (2009). Contribuição de Ferramenta e Práticas de Gestão da Qualidade, Tecnologia e Conhecimento para a Evolução do Nível de Maturidade do processo de Distribuição de Energia Elétrica no Brasil – A Percepção do Cliente e do Gestor do Processo. Master Dissertation. Universidade Tecnológica Federal do Paraná – UTFPR, Paraná, Brazil.

[13] L. E. L. Gonçalves, 'As Empresas São Grandes Coleções de Processos'. Revista de Administração de Empresas. Jan/Mar 2000. São Paulo v. 40, n.1, pp. 6-19.

[14] N. Kagan, C. C. B. Oliveira, and C. C. B. de Guaraldo, (2000) – Enerq /USP J. S. Simões – CELESC. Planejamento de redes de distribuição a partir de avaliação automática de reforços na rede.

[15] N. Kagan, (1992) Electrical Power Distribution Systems Planning Using Multiobjective and Fuzzy Mathematical Programming. PhD thesis. Queen Mary & Westfield College, University of London, United Kingdom.

[16] N. Kagan, (1999) Configuração de Redes de Distribuição Através de Algoritmos Genéticos e Tomada de Decisão Fuzzy. Tese de Livre Docência: Escola Politécnica da Universidade de São Paulo – EPUSP, São Paulo, Brazil.

[17] N. Kagan, et al. (2004). INTERPLAN – Uma Ferramenta para o Planejamento Integrado de Sistemas de Distribuição de Alta, Média e Baixa Tensão. IEEE T&D Latin America, São Paulo.

[18] K. C. Laudon and L.P. Laudon, (2004) Management Information System: managing the Digital Firm. 8. Ed., New Jersey: Ed. Prentice-Hall.

[19] I. Nonaka and H. Takeuchi, (1997). Criação do Conhecimento na Empresa. Rio de Janeiro: Campus.

[20] C. C. B. Oliveira, (1997) Configuração de Redes de Distribuição de Energia Elétrica com Múltiplos Objetivos e Incertezas Através de Procedimentos Heurísticos. . PhD thesis. Escola Politécnica da Universidade de São Paulo – EPUSP, São Paulo, Brazil.

[21] Relatório Plano de Desenvolvimento de Distribuição – (PDD) Fonte Celesc/ DPEP/DVPE, 2009.

[22] M. R. Rodrigues and L. Helena, Um Modelo de Gestão do Conhecimento em uma Empresa de Energia, II Simpósio Internacional de Transparência nos Negócios, 31 Julho a 2 de Agosto de 2008.

[23] E. Santana, Em entrevista para Canal Energia, Blecautes, regulação e investimentos. Obtained Through the Internet 2010.http://www.canalenergia.com.br/spublisher/materiais/Artigos_e_Entrevistas.asp?id=76862. (retrieved: November, 2011).

[24] P. M. Senge, et al. (1994). A quinta disciplina – caderno de campo: estratégias para construir uma organização que aprende. Rio de Janeiro: Qualitymark.

[25] S. L. Silva, 'Gestão do conhecimento: uma revisão crítica orientada pela abordagem da criação do conhecimento'. Ciência da Informação. Brasília, V.33, n.2, pp. 143-151, may-august 2004.

[26] J. Stecjuka, J. Makna and M. Kirikova. Best practices oriented business process operation and design. pp. 122-129. Proceedings of BPMDS' 2008.

[27] J. Reichwald, T. Dornemann, T. Barth, M. Grauer and B. Freislebem. Model-Driven Process Developent Human Task in Service Grid Environments. pp. 79-90.

 ISBN: 978-1-61208-181-6

Enhancing Bayesian Network Model for Integrated Software Quality Prediction

Łukasz Radliński
Institute of Information Technology in Management
University of Szczecin
Szczecin, Poland
lukrad@uoo.univ.szczecin.pl

***Abstract*—A Bayesian network model for integrated software quality prediction, proposed in earlier study, has potential in supporting decision makers in software projects. However, it also has some disadvantages limiting its use. The aim of this paper is to overcome these limitations by enhancing the original model in three ways: (1) incorporating project factors, (2) adding subnets with detailed process factors, and (3) modeling integration of software components or sub-systems. These enhancements significantly improve the analytical usefulness of this predictive Bayesian network model.**

Keywords-Bayesian network; decision support; process factors; project factors; quality factors; software quality.

I. INTRODUCTION

The quality of software is a very important aspect of software project. Thus, software quality have been extensively studied since the turn of 1960's and 1970's [1][26]. While most of these studies have been focused on software defectiveness [6], some researchers also investigate selected features of software quality such as reliability [16][18], maintainability [25] or usability [2]. Although such studies are very useful contributions to software engineering discipline, they typically focus on a single feature of software quality.

Project decisions related to software quality require support from analytical and predictive models. It is possible to make decisions based on output from models focusing on a single quality feature. The majority of existing approaches involving techniques, such as case-based reasoning, decision trees, multiple regression, are not feasible for this purpose because they focus on a single output. Important decisions, influencing the whole project and its environment, should be made after deeper analyses of possible effects involving multiple outputs. Performing such analyses can be supported by a simulation model that can handle multiple outputs and various types of relationships. In our experiences with using empirical data in software companies, we found that the companies do not have data of required volume and granularity to automatically generate/learn the model purely from data. Therefore, we propose using expert-driven Bayesian networks (BNs) as a formal representation for such simulation model. Section III provides more details on motivations for using BNs.

Earlier studies [20][21] proposed a BN model for integrated software quality prediction. Preliminary experiments revealed that this model may be a useful simulation tool for decision makers in software projects. The main aim of this paper is to develop an enhanced version of this predictive model. The main contributions of this paper are the following enhancements of the original model:

- Incorporating project factors that describe the nature of a project – as a result, an enhanced model can be reused for different types of software projects, rather than for a single project type defined upfront;
- Adding subnets with detailed process factors influencing overall process quality – this may be useful where direct assessment of the level of process quality is difficult or where it is useful to perform simulations using detailed process factors;
- Modeling an integration of software components or sub-systems into larger software products – this extends the usability of the model for different parts of a software product and their integration.

This paper is organized as follows: Section II defines software quality and its factors according to ISO standards. Section III discusses related work. Section IV summarizes original BN model. Section V presents proposed enhancements to the original model. Section VI provides plans for model calibration and validation. Section VII draws conclusions and discusses future work.

II. SOFTWARE QUALITY FACTORS

Detailed analysis of software quality requires investigating a variety of quality factors. This paper is based on the breakdown of software quality proposed in ISO 250xx series of standards [11][12], which superseded an older 9126 standard [13]. On the first level there are 11 quality features: compatibility, flexibility, functional suitability, maintainability, operability, performance efficiency, portability, reliability, safety, security, and usability. Then each feature is decomposed into a set of sub-features. For example, reliability has five sub-features defined: availability, fault tolerance, maturity, recoverability, and reliability compliance. On the third level there are measures describing specific sub-features. These measures should be carefully selected depending on the purpose of analysis and environment where such model will be used. In this paper, a term 'quality factors' refers to all levels of software quality, i.e., features, sub-features and measures.

 ISBN: 978-1-61208-181-6

III. RELATED WORK

In previous research, a variety of statistical and machine learning techniques have been used for quality prediction. The most popular are: multiple regression (MR), case-based reasoning (CBR), decision trees (DT), random forests (RF), rule induction (RI), support vector machines (SVM), system dynamics (SD), neural networks (NN), and Bayesian networks (BN). We have investigated various features of popular and well established techniques. This analysis helped in selecting the technique that would be the best suited for our model for software quality prediction.

Table I illustrates how various features of modeling, simulation and prediction correspond to different techniques. This comparison has been developed based on the extensive literature survey, involving the investigation of inherent features of these techniques [17], applications of these techniques in software engineering area [5][7][28][30][32], our own experiments – both published [24] and unpublished. With this comparison we do not attempt to produce a general ranking of techniques, since it is very difficult and probably not possible [14][23] or feasible [29], because the technique selection should involve context-specific features. In this comparison we do not consider the accuracy of predictions for these techniques. Earlier studies showed that the accuracy is varying significantly depending on particular dataset used in analysis [14][17][24][28][32].

Most of these techniques are data-driven, which means that the prediction is provided almost entirely based on empirical data. Thus, these techniques fail when such data is not available. We were aware that, due to availability of the data, our model would have to be based in larger extent on expert knowledge rather than on empirical data.

Additionally, only some of these techniques enable providing prediction for multiple dependent variables. Such functionality is crucial because we attempt to develop a model where software quality is reflected not by a single variable but a range of interrelated variables.

The main use of the model is to provide decision support through the ability of performing various simulations. To make these simulations more realistic, the model should have the ability of defining causal relationships by domain experts. Only very few techniques enable this feature.

TABLE I. FEATURES OF POPULAR MODELLING TECHNIQUES[a]

Feature	Technique								
	MR	***CBR***	***DT***	***RF***	***RI***	***SVM***	***SD***	***NN***	***BN***
expert knowledge	L	H	H	L	H	L	H	M	H
multiple dependent variables	L	L	L	L	H	L	H	H	H
causal relationships	M	L	M	M	H	L	H	M	H
explit uncertainty	M	L	L	L	L	L	L	L	H
intuitiveness	H	M	H	L	H	L	H	L	H
ease of adaptation	H	M	M	M	M	H	M	M	M

a. 'L' – low, 'M' – medium, 'H' – high

Given the context of our research, we have selected BN as a formal representation for our predictive model, because this technique enables the required functionality. BN is a very powerful modeling technique that has already been widely used in various studies on software engineering [22]. Pfautz at al. say that they are "well-suited to capturing vague and uncertain knowledge" [19]. BNs have a unique set of features such as ability to incorporate expert knowledge and empirical data, explicit modeling of causal relationships, probabilistic definition of variables reflecting uncertainty of modeled system, no need to declare in advance a list of input and output variables, ability to run with incomplete data, and visual representation. BNs can also take a form of time-series models called dynamic Bayesian networks. Detailed analysis of the motivations for using BNs can be found in [8][23].

BNs have been used in earlier studies to model software quality. However, most of these studies have been focused on a single aspect of software quality. We found three references, where the authors model multiple features of software quality.

Beaver [4] developed a BN model to reflect software quality according to the ISO 9126 standard. However, the author does not provide enough details on model structure, variable definitions and model validation. Thus, it is difficult to assess the correctness and usability of this model.

Wagner [31] proposed BN models for predicting software quality using activity-based software quality models. In contrast with the current study, the author focused on modeling selected features from ISO 9126 standard, i.e., maintainability and security, and not the relationships between these features.

Fenton at al. [10] developed a BN model for the trade-off between development effort, project scope and software quality. In this model software quality is reflected by two variables, defect rate and customer satisfaction.

IV. ORIGINAL BAYESIAN NETWORK MODEL

The main aim of the BN model is to deliver useful information to project managers and support their decisions. Proposed BN model enables performing various types of analyses:

- 'What-if' analysis - investigating how different actions may influence specific quality factors. For example, an impact of increased amount of *specification effort* on *functional suitability*, *maintainability* or *operability*.
- 'Goal-seeking' - answering a question: how to achieve a specific target? For example: how much better a testing process is required to achieve a higher level of reliability (with other constraints entered to the model).
- 'Trade-off' analysis - investigating the degree at which a quality factor that has to be traded for another quality factor (given other constrains). For example: an architectural trade-off between *performance efficiency* and *maintainability*, where efficient software may be difficult in maintenance.

The initial structure of this model has been discussed in [20][21]. This model is too large to be presented in detail here. Thus, this paper only briefly summarizes its main concepts illustrated in Figures 1 and 2. The main parts, i.e., quality factors are modeled as hierarchical Naïve Bayesian Classifiers where variables reflecting a detailed level of software quality are the children of the more general factors. For example, usability has four sub-features modeled as its children (Figure 1).

Such structure enables easy adjustments, e.g. adding a new sub-feature requires only a definition of this newly added variable without a need to change other parts of the model. Such structure works well even when relationships between children variables exist in reality but have not been included in the model [27].

Quality features are linked with other. These links have been defined according to a knowledge base that contains results from a literature survey [20][21]. Figure 2 illustrates some of these links.

The model also contains some basic process variables describing *effort*, *process quality* and *process effectiveness*. Since this is a part of the model that was significantly enhanced more details on process variables have been provided in Section V.A.

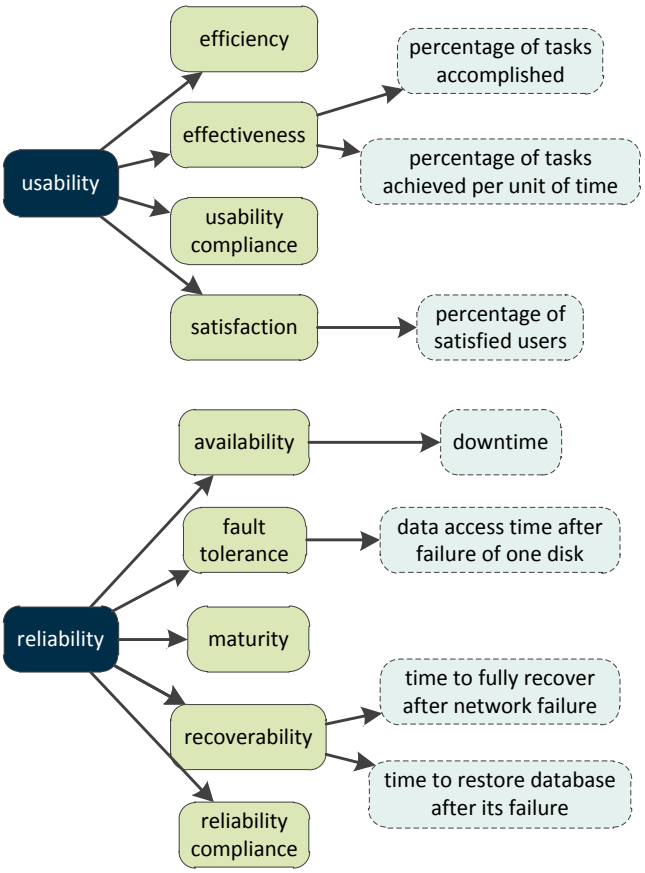


Figure 1. Three levels of software quality: features (left), sub-features (center) and examples of measures (right).

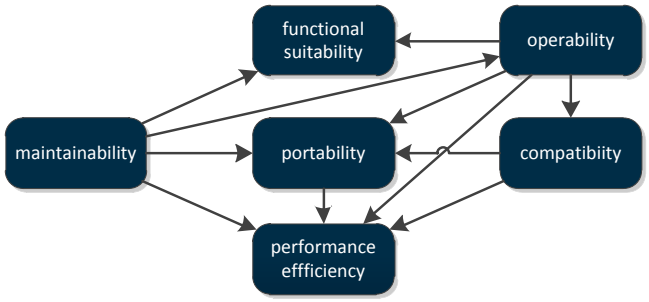


Figure 2. Selected links between quality features.

Proper definition of the quantitative part, i.e., probability distributions for variables, is a challenging step in the process of building a BN. All variables in this model are defined using a 5-point ranked scale from 'very low' to 'very high' level of intensity. Probability distributions are not defined by manually filling probability tables but using a set of expressions such as weighted mean, weighted max and weighted min [9]. For example, the following expression:

$$proc\_eff = \mathrm{N}(\mathrm{wmean}(3, effort, 4, process\_q), 0.001) \quad (1)$$

means that *process effectiveness* is defined by a Normal distribution as a weighted mean of *effort* and *process quality* with weights 3 and 4, respectively; 0.001 is a value of variance and represents the level of uncertainty. Such types of expressions simplify the process of building a BN because they require only the values of the weights for each variable instead of the whole probability tables.

V. MODEL ENHANCEMENTS

This section considers three main enhancements of the original BN model: incorporating project factors (Subsection A), adding subnets with detailed process factors (Subsection B) and integrating software components or sub-systems (Subsection C).

A. Project Factors

Original model did not contain any project factors, i.e., factors describing the nature of developed project. Thus, it had to be calibrated separately for each project or, more generally, for each type of project. Since such calibration is time-consuming, to improve model usefulness the enhanced model contains additional project factors. These project factors reflect the nature of the project and its environment. Currently the model contains the following project factors: *architecture*, *CASE tool used*, *development platform*, *functional size*, *UI* (User Interface) *type*, *intended market*, and *used methodology*.

Figure 3 illustrates links between selected project factors and selected quality features. To simplify the definition of probability tables for quality features the model uses so called 'partitioned expressions', where the child node is defined using different expressions for different states of parent nodes. Figure 4 provides an example of such expressions for *operability* given selected states of *UI type,* together with visualization of probability distributions.

 ISBN: 978-1-61208-181-6

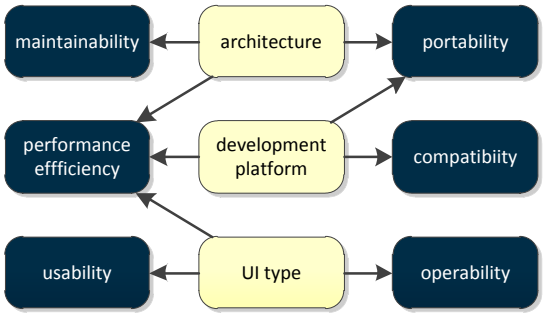


Figure 3. Example of modelling the impact of project factors.

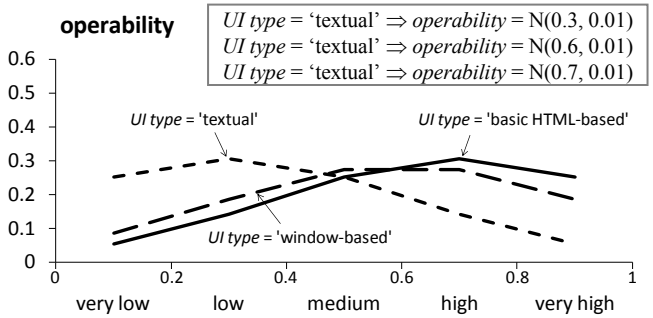


Figure 4. Example of modelling the impact of project factors.

A difficulty of this enhancement is related with the calibration stage. Some quality features, such as *performance efficiency*, have many project factors as parents. Defining probability distributions for such quality features is difficult because they have to reflect every possible combination of states of parent project factors. However, after performing such calibration the usability of the BN model is significantly improved.

B. Process Factors

The original model contains nine variables reflecting process of software development: *effort*, *process quality* and *process effectiveness* – separately for three main activities of software development: specification, implementation and testing. In some situations it might be sufficient to represent process quality as a single variable. However, to improve the analytical capabilities of this model, it has been enhanced by subnets with detailed process factors, separate for three main activities of software development.

Figure 5 illustrates a subnet for process factors in specification. Subnets for implementation and testing stages have similar structures – the difference is that they do not contain variables related to requirements (upper right part of Figure 5). Variables describing process factors have been mainly linked according to causal relationships.

For example, the level of *leadership quality* influences three variables: *team organization*, *defined process followed* and *appropriateness of methods and tools used*. Then, the level (quality) of *requirements management*, *defined process followed* and *appropriateness of methods and tools used* jointly determine the level of *process quality*. *Process quality*, *requirements creep* and *staff quality* influence the *overall process quality*.

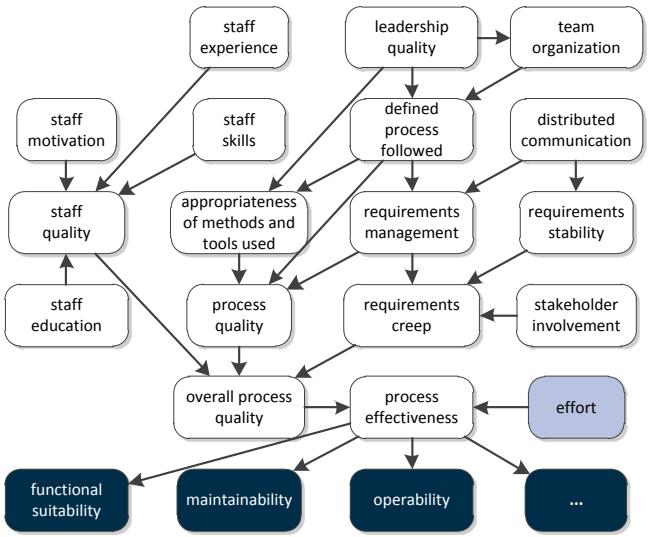


Figure 5. Subnet for process factors in specification stage.

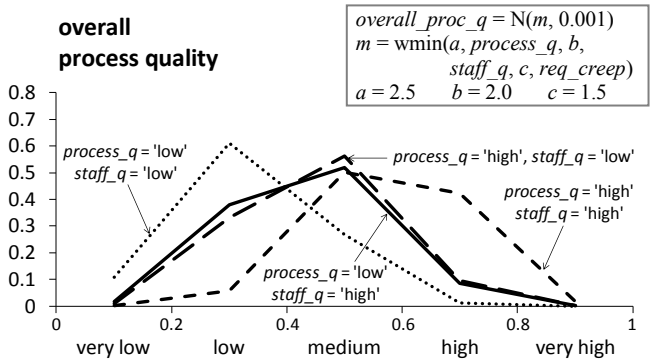


Figure 6. Example of aggregation of process factors.

Process effectiveness aggregates all process factors and is defined as a combination of *overall process quality* and the level of *effort*. Finally, *process effectiveness* variables, separately in three development stages, influence the level of selected software quality features. Variables in this subnet are quantitatively defined using weighted expressions similar to Equation 1.

Figure 6 illustrates the impact of various combinations of *process quality* and *staff quality* on *overall process quality*. The latter is defined as a weighted min (wmin) of its parents to incorporate the fact that undesirable state of one parent node may significantly decrease the value of *overall process quality*, even if other parents are at desirable states. The values of weights determine the strength of impact of particular parent on the aggregated value.

C. Integrating components/sub-systems

One of the challenges of building a predictive model for software quality is to properly define the level of granularity. Such model may be built for the whole software systems, sub-systems, single applications, components, modules, classes etc. To improve the flexibility of this model, as another enhancement of the original model, it now can be used at various levels of details.

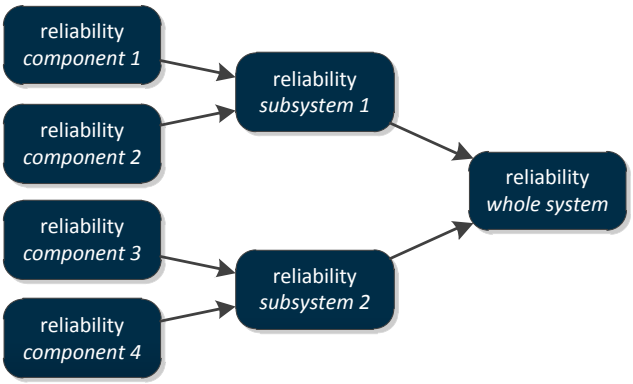


Figure 7. Example of integrating components/sub-systems.

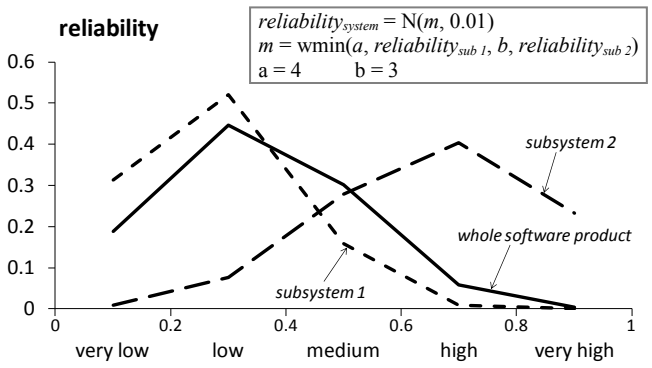


Figure 8. Definition and predictions for integrated reliability.

The basic idea, illustrated in Figure 7, is the following: First, users nominate the lowest level of details for the model. Then, they perform a calibration for this level. The higher level of details is modeled by aggregating quality factors (and possible other variables too). This aggregation can be done using expressions such as mean, max, min, weighted mean, weighted max or weighted min [9]. This procedure may be continued until the highest desired level of details has been reflected in the model.

This enhancement enables performing predictions for quality factors either at the detailed level, e.g., a class or a module, or aggregated, e.g., an application or a system. An example of such aggregation for reliability is shown in Figure 8. The reliability of the whole software product can only be as high as for the part with the lowest level of reliability. In the model it is reflected with the weighted min (wmin) function. Because some subsystems may be more frequently used than the other, their impact on overall reliability would be greater. This can be reflected by adjusting the values of weights *a* and *b*. In this hypothetical example a subsystem 1 is more reliable than subsystem 2. It also has greater impact on overall reliability (a>b). Therefore, the reliability of the whole system is between the level of reliability for subsystems 1 and 2, but much closer to reliability of system 1.

VI. PLANS FOR MODEL CALIBRATION AND VALIDATION

Currently, the process of model calibration with industrial partners is performed. In the first step, a tradeoff between model complexity, usability and clearness is investigated. It is focused on answering a question: how large the model can be so that it is clear enough for being used in industry? This calibration is performed using a customized technique of structured interviews based on repertory grid [3][15].

After investigating some patterns from this analysis, the model structure will be adjusted. Then, detailed model calibration will be performed using structured interviews. This will enable capturing relevant expert knowledge that would be difficult to express using only a predefined questionnaire.

This calibration will cover almost the whole structure of the model – except the quantitative measures assigned to quality features/sub-features. Companies that accepted to participate in the process of calibrating the model are not willing to provide such data outside their environments. On one side this is related with data protection and privacy, on the other side with time consuming process of preparing them. Calibration of the rest of the model will be performed by asking domain experts to:

- Assign weights in the weighted expressions;
- Assign the level of their uncertainty about provided data;
- Provide prior distributions for root nodes.

Results of this survey will be combined with results available in the literature and empirical analyses performed earlier.

The internal validation of the model will be focused on investigating how well the model incorporates data/knowledge gathered during the calibration stage. A variety of fitness measures will be used here. In the external validation, industrial partners will be granted access to the model to familiarize with it and assess a variety of its features, such as correctness, usability, clearness, ease of use and ease of customization/calibration.

VII. CONCLUSIONS AND FUTURE WORK

Proposed BN model for software quality prediction reflects the breakdown of quality factors proposed in ISO 250xx series of standards. It contains a variety of software quality factors, together with relationships between them. It also contains process factors that influence software quality.

Obtained results lead to the following conclusions:

- Original BN model is useful in a variety of applications but suffers limitations related to the lack of details on selected software development aspects;
- Proposed enhancements, i.e., incorporating project factors, adding subnets with detailed process factors and ability of integrating software components or sub-systems overcome these limitations;
- Proposed enhancements require additional time for model calibration in target environment.

 ISBN: 978-1-61208-181-6

Plans for future work related to this BN model include:

- Further enhancements to the model to reflect the dynamics of software development and maintenance;
- Automated calibration of the model using the data from software repositories or knowledge base;
- Detailed model calibration and validation using software engineering literature, expert judgment, and empirical data.

ACKNOWLEDGMENT

This work has been partially supported by research funds from the Ministry of Science and Higher Education in Poland as a research grant no. N N111 291738 for years 2010-2012.

REFERENCES

[1] F. Akiyama, "An Example of Software System Debugging," in Proceedings of Federation for Information Processing Congress, vol. 71, Ljubljana, 1971, pp. 353-379.

[2] A. Abran, A. Khelifi, W. Suryn, and A. Seffah, "Usability Meanings and Interpretations in ISO Standards," Software Quality Journal, vol. 11, pp. 325-338, 2003.

[3] H. C. Banestad, J. E. Hannay, "Comparison of Model-based and Judgment-based Release Planning in Incremental Software Projects," in Proceeding of the 33rd International Conference on Software Engineering, ACM, 2011, pp. 766-775.

[4] J. M. Beaver, "A life cycle software quality model using bayesian belief networks," Ph.D. Thesis, University of Central Florida, 2006.

[5] S. Bouktif, F. Ahmed, I. Khalil and G. Antoniol, "A novel composite model approach to improve software quality prediction," Information and Software Technology, vol. 52, no. 12, pp. 1298-1311, Dec. 2010.

[6] C. Catal and B. Diri, "A systematic review of software fault prediction studies," Expert Systems with Applications, vol. 36, no. 4, pp. 7346-7354, May. 2009.

[7] C. Catal, "Review: Software fault prediction: A literature review and current trends", Expert Systems with Applications, vol. 38, no. 4, pp. 4626-4636, Apr. 2011.

[8] N. E. Fenton and M. Neil, "A Critique of Software Defect Prediction Models", IEEE Transactions on Software Engineering, vol. 25, no. 5, pp. 675-689, Sep. 1999.

[9] N. E. Fenton, M. Neil, and J. G. Caballero, "Using Ranked Nodes to Model Qualitative Judgments in Bayesian Networks," IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 10, pp. 1420-1432, Oct. 2007.

[10] N. Fenton, W. Marsh, M. Neil, P. Cates, S. Forey, and M. Tailor, "Making Resource Decisions for Software Projects," in Proceedings of the 26th International Conference on Software Engineering, 2004, pp. 397-406.

[11] ISO/IEC 25000:2005, Software Engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Guide to SQuaRE, 2005.

[12] ISO/IEC CD 25010:2008, Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Software and quality in use models, Version 0.55, 2008.

[13] ISO/IEC FDIS 9126-1:2001, Software Engineering – Product quality – Part 1: Quality model, 2001.

[14] Y. Jiang, B. Cukic and T. Menzies, "Cost Curve Evaluation of Fault Prediction Models," in Proceedings of the 2008 19th International Symposium on Software Reliability Engineering, IEEE Computer Society, Washington, DC, 2008, pp. 197-206.

[15] G. Kelly, "The psychology of personal constructs," Norton, New York, 1955.

[16] M. Lyu, "Handbook of software reliability engineering," McGraw-Hill, Hightstown, NJ, 1996.

[17] C. Mair, G. Kadoda G, M. Lefley, K. Phalp, C. Schofield, M. Shepperd and S. Webster, "An investigation of machine learning based prediction systems," Journal of Systems and Software, vol. 53, no. 1, pp. 23-29, Jul. 2000.

[18] J. D. Musa, "Software Reliability Engineering: More Reliable Software Faster and Cheaper," Second Edition, Authorhouse, 2004.

[19] J. Pfautz, D. Koelle, E. Carlson, and E. Roth, "Complexities and Challenges in the Use of Bayesian Belief Networks: Informing the Design of Causal Influence Models," Human Factors and Ergonomics Society Annual Meeting Proceedings, vol. 53, no. 4, pp. 237-241, Oct. 2009.

[20] Ł. Radliński, "A Conceptual Bayesian Net Model for Integrated Software Quality Prediction," Annales UMCS Informatica, vol. 11, no. 2, 2011 (accepted).

[21] Ł. Radliński, "A Framework for Integrated Software Quality Prediction using Bayesian Nets," in Proceedings of International Conference on Computational Science and Its Applications (ICCSA 2011), vol. 6786, Springer, 2011, pp. 310-325.

[22] L. Radlinski, "A Survey of Bayesian Net Models for Software Development Effort Prediction", International Journal of Software Engineering and Computing, vol. 2, no. 2, pp. 95-109, 2010.

[23] Ł. Radliński, "Techniques for Predicting Development Effort and Software Quality in IT Projects", Research Papers of the University of Szczecin. Series: Studia Informatica, vol. 26, pp. 119-137, 2010 (in Polish).

[24] L. Radlinski and W. Hoffmann, "On Predicting Software Development Effort using Machine Learning Techniques and Local Data", International Journal of Software Engineering and Computing, vol. 2, no. 2, pp. 123-136, 2010.

[25] M. Riaz, E. Mendes, and E. Tempero, "A systematic review of software maintainability prediction and metrics," in Empirical Software Engineering and Measurement, 2009, pp. 367-377.

[26] R. J. Rubey, R. D. Hartwick, "Quantitative measurement of program quality," in: Proceedings of ACM National Conference, ACM, 1968, pp. 671-677.

[27] S. Russell, P. Norvig, "Artificial Inteligence. A Modern Approach," Second Edition, Pearson Education, Upper Saddle River, 2003.

[28] M. Shepperd and G. Kadoda, "Comparing Software Prediction Techniques Using Simulation," IEEE Transactions on Software Engineering, vol. 27, no. 11, pp. 1014-1022, Nov. 2001.

[29] Q. Song, Z. Jia, Shepperd M., S. Ying and J. Liu, "A General Software Defect-Proneness Prediction Framework", IEEE Transactions on Software Engineering, vol. 37, no. 3, pp. 356-370, May-Jun. 2011.

[30] B. Stewart, "Predicting project delivery rates using the Naive–Bayes classifier", Journal on Software Maintenance and Evolution: Research and Practice, vol. 14, pp. 161-179, 2002.

[31] S. Wagner, "A Bayesian network approach to assess and predict software quality using activity-based quality models," Information and Software Technology, vol. 52, no. 11, pp. 1230-1241, Nov. 2010.

[32] D. Zhang and J. J. P. Tsai, "Machine Learning and Software Engineering," Software Quality Journal, vol. 11, no. 2, pp. 87-119, 2003.

 ISBN: 978-1-61208-181-6

Mapping OBI and XPDL to a MDE Framework for Laboratory Information Processing

Alessandro Maccagnan*, Nicola Cannata‡, Giorgio Valle†, Tullio Vardanega*
*Department of Pure and Applied Mathematics, University of Padua, via Trieste 63, 35121 Padova, Italy
e-mail:{maccagnan, tullio.vardanega}@math.unipd.it
‡School of Science and Technology, University of Camerino, Via Madonna delle Carceri 9, 62032 Camerino, Italy
e-mail: nicola.cannata@unicam.it
†CRIBI Biotechnology Centre, University of Padua, viale G. Colombo 3, 35121 Padova, Italy
e-mail: giorgio.valle@cribi.unipd.it

***Abstract*—Biomedical analyses are becoming increasingly complex, both for the type of data produced and the procedures necessary to obtain them. This trend is expected to continue; therefore the development of suitable systems for information and protocol management is becoming essential for the full exploitation of the field. Custom-built applications obtained by direct merging of software engineering expertise with domain-specific knowledge may be temporary solutions, but they are generally ineffective both in terms of cost and performance. Here we propose a Laboratory Information Management System (LIMS) that enables the domain experts to express laboratory protocols using domain knowledge, free from the incidence and mediation of the software implementation artifacts. In the system that we propose this is made possible by basing the modeling language on an authoritative domain specific ontology.**

***Index Terms*—Model-Driven Engineering; Laboratory Protocols; Ontology; Process Definition Language.**

I. Introduction

A. Motivation and Vision

In the last two decades life sciences and biomedicines have been revolutionized by the introduction of high-throughput procedures and automation methods. Laboratory Information Management Systems (LIMS) are tools used for tracking protocols and samples, in order to reliably cope with such turnout. Unfortunately protocols are still mainly written and exchanged in natural languages which is a serious impediment to quality, efficiency, predictability and repeatability. We claim we should rather strive to represent protocols in a structured and efficient way. In this work, we present a way to bridge the benefits of Ontology onto Model-Driven Engineering (MDE) in order to satisfy this need.

What we ultimately aim to accomplish is to tie in an intimate way ontologies and meta-models. Our aim is to build models that are deeply-rooted on ontologies. We propose a conceptual framework that links a *construct that describes reality* (ontology) to a *construct that prescribes reality* (model). Hence we are trying to bind ontological constraints directly to model elements.

The structure of the paper is as follows. In section II, we give a brief overview on Model-Driven Engineering and Ontology and some relevant literature on how to merge them. In section III, we describe our proposal on how to merge a domain specific ontology with a workflow metamodel.

II. Background

A. Model-Driven Engineering

MDE is an approach to software development which concentrates on designing models that are closer to domain-specific concepts of some particular domain rather than to computing (or algorithmic) ones. MDE's basic concepts are models, meta-models and transformations [1].

A model is a "set of statements about some system under study" [2]. In traditional scientific disciplines, models are usually descriptive. However they are also used as specifications in engineering disciplines, including software design. Therefore a model could equally be descriptive or prescriptive.

A distinctive trait of models is their intended relationship with reality: "A model is an external and explicit representation of a part of reality as seen by the people who wish to use that model to understand, change, manage, and control that part of reality" [3].

Models can represent, describe, and specify things [4]. A *descriptive* model is one that "describes reality, but reality is not constructed from it". A *prescriptive* model is one that "prescribes the structure or behavior of reality and reality is constructed according to the model; that is, the model is a specification for reality" [2]. Since in the realm of software engineering most of the models are used to construct a "reality" from them, in the remainder of this paper we understand a model as prescriptive.

B. Ontologies

The term 'ontology' is currently very controversial controversial because different people have different ideas on the definition of an ontology. However there is a certain consensus in what an ontology is not: it is neither a taxonomy (i.e., a class-subclass hierarchy), nor a dictionary (an ontology does include relationships between terms), nor a knowledge base that includes only individual objects. According to Gruber, an ontology can be defined as "the specification of conceptualizations, used to help programs and humans share knowledge" [5].

 ISBN: 978-1-61208-181-6

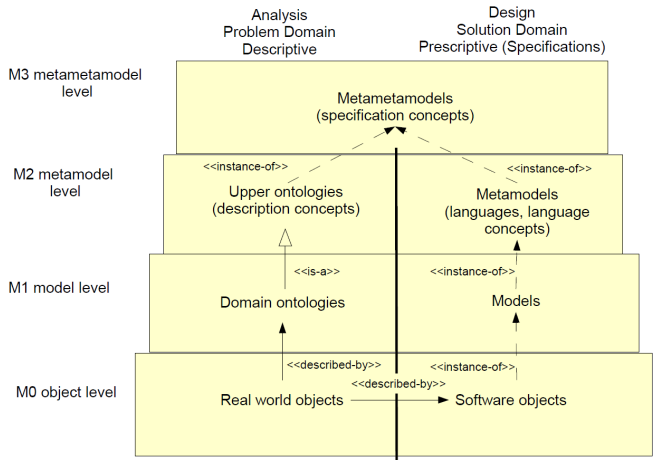


Fig. 1. The ontology-aware meta-pyramid. Domain ontologies live at level M1. Upper ontologies live at level M2. Ontology metalanguages live at level M3. Source: [4].

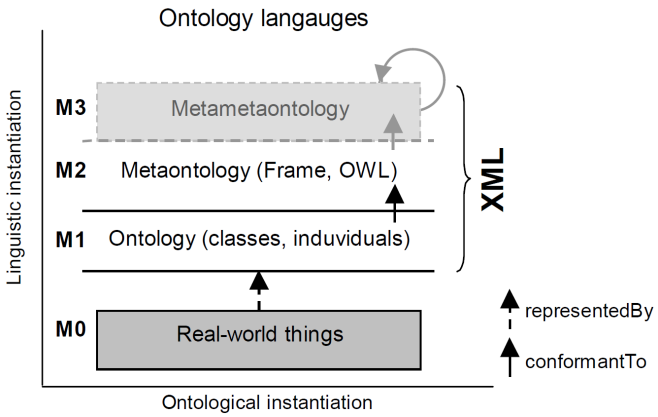


Fig. 2. The four meta-layers in terms of ontological engineering and its orthogonal instance-of relations: linguistic and ontological. Source: [8].

The first formal and explicit approach to ontologies in the technical (not philosophical) sense dates back to 1900, given by Husserl. Later in the 1980's, the ontologies entered the computer science field as a way to provide a simplified and well defined view of a specific area of interest or domain. Semantic web is the modern expression of the field. The Web Ontology Language (OWL) is a modern ontological language endorsed by the World Wide Web Consortium (W3C).

What we really are after is the conceptual relationship between a model and an ontology in the context of knowledge and process management.

An important property of ontologies is the *open-world assumption*, i.e., anything not expressed is unknown [6]. In models the *closed-world assumption* is generally used since what has not been specified is not unknown but true (or false) by default. As we noted earlier, models are usually prescriptive tools. What can we say for ontologies? Ontologies are not specification models since they describe domains and not systems [4].

Ontologies are tools extensively used to express domain knowledge. One serious problem is that differing ontologies may be developed and applied for the representation of one and the same domain. The function of an upper ontology is precisely to "support interoperability between domain ontologies in order to facilitate the shared use of data both within and across disciplinary boundaries" [7]. A domain ontology specializes concepts taken from an upper ontology.

C. MDE and Ontologies

Assmann et al. propose the ontology-aware meta-pyramid [4] (Fig. 1) in order to show how ontologies can be used in MDE. Domain ontologies live at level M1 of the meta-pyramid and correspond to models. An upper ontology, providing a language for ontologies, should live at level M2. One metametamodel language (at level M3) could be used to specify both ontology and metamodels. Both the ontology dimension and the model-driven dimension instantiates from this metametamodel.

Bezivin et al. use a different approach to relate ontology and MDE. The four MDE levels are called in this context linguistic layers [8]. Concepts from the same linguistic layer can be at different ontological layers. Figure 2 depicts the four meta-layers using this important remark. The linguistic instantiation runs on the vertical dimension; conversely the ontological instantiation runs on the horizontal dimension (e.g., like for an upper ontology and a domain ontology built upon it).

Kapsammer et al. propose a mapping between the metametamodel language Ecore (model engineering space) and the ontology definition metamodel (ontology technical space) [9]. Ecore belongs to the Eclipse Modeling Framework (EMF) and it is used to describe models and metamodels. Also the mapping proposed between EMF objects and OWL/RDF resources [10] presents some difficulties (e.g., class membership), because in object-oriented languages the membership of objects is fixed along a derivation hierarchy. In OWL instead, individuals can belong to multiple classes.

Hillairet et al. designed a set of Eclipse plugins that are able to make a round-trip transformation between OWL and Ecore. The project (named EMF4SW) is not yet mature enough to cope with large and complex ontologies. However it is in very active development and it is able to deal with relative small ontologies.

Parreiras et al. describe a vision in which both worlds (MDE and Ontology) co-exist under a common umbrella [11]. The concepts of Metamodeling Technical Spaces (MMTS) as well as Ontological Technical Spaces (OTSs) are introduced, derived from the work of [12]. They suggest some strategies for integrating OTS into MMTS. Furthermore they provide a list of desirable features for the "marriage" MMTS+OTS.

III. Combining Ontologies and Workflows

Our field of application is that of laboratory informatics and scientific experimentation. Relying on our experience and on related literature we observed that laboratory procedures are based on some key elements ([13], [14]). They consist

 ISBN: 978-1-61208-181-6

in workflows that describe some interaction between some objects by means of some actions. As far as the descriptive knowledge of this domain is concerned, the most relevant ontology accepted by the community is the Ontology for Biomedical Investigations (OBI). For the prescriptive part of our effort instead, we opted for on BPMN/XPDL for workflows definition.

A. Ontology for Biomedical Investigations

OBI is an ontology for the description of biological and clinical investigations [15]. OBI describes the design of an investigation, protocols and instrumentation, materials used, data generated and analysis performed on it. The OBI project is developed in the frame of the Ontology for Biomedical Investigations (OBO) Foundry [16], and as such, it adheres to the principles of OBO, as orthogonal coverage and the use of a formal language. OWL was chosen as the OBI language. The ontology is developed to model biomedical investigations, therefore it contains terms for aspects such as:

- Biological material, e.g., DNA
- Instrument, e.g., centrifuge or thermal cycler
- Design and execution of an investigation, e.g., injecting mice with a vaccine to test its efficacy.

OBI relies on the Basic Formal Ontology (BFO) upper ontology. An upper ontology describes concepts of the "Reality" from a high-level of abstraction.

B. XML Process Definition Language

XPDL is a markup language created to ensure interoperability among different workflow management systems. Its main goal is to exchange process definitions, addressing both the graphical and the semantic notations.

The meta-model of XPDL involves the definition of activities, the specification of their order of execution and the involved data. The flow of execution is specified through different constructors: sequence, split, join. An elementary activity is an atomic piece of work [17]. An Activity could modify relevant data declared as DataField. In addition to standard types a user could add external types (by means of an XSD declaration or an external reference).

C. Mapping XPDL and OBI

Finding a method to relate the MDE architecture (its layers of abstraction) to the ontology schema is key to facilitating the systematic use of ontologies inside prescriptive models. Based on the literature we reviewed we built a relation between the classical layers of the MDE architecture, BFO/OBI and XPDL.

Fig. 3 depicts the classic layers of MDE. The workflow components of our formalism are fairly easy to place within this hierarchy. XSD, the XML schema language used to describe XPDL, can be positioned at the M3 level (i.e., meta-meta-model). XPDL conforms to a XSD model and therefore lies at the M2 level (i.e., meta-model). A valid XPDL workflow (i.e., a model for the end user) is at level M1. A specific esecution of a workflow resides at the ground level M0 (not shown in the figure).

In XPDL, the concept of Activity represents the unit of work. An Application is a particular kind of Activity that describes functionalities offered by legacy systems. In XPDL an Application is invoked by means of a Tool Activity. In object-orientation terms, an Application can be seen as an interface for a functionality with a name and a list of parameters. We can think of an interface as a sort of "contract" between a class and the outside world. Every parameter is described with a name, a type, and a mode of passing (input, output, mixed). The Application construct represents the junction point between the workflow world of XPDL and the ontological world of BFO/OBI.

Before defining a mapping between BFO/OBI and XPDL we need to also relate the former to MDE. BFO is written using OWL, hence, in our schema of interpretation, OWL is at level M3 and BFO at M2. OBI is a specialization of BFO in the dimension of the description of the domain. It is not a specialization in the linguistic dimension proper of the MDE [4]. For that reason OBI places at M2 but in a sort of orthogonal dimension to the classic hierarchy (which we show horizontally instead of vertically in Fig. 3). A consequence of this is that instances of BFO/OBI concepts (in OWL called individuals) are at M1. Using this schema of interpretation individuals are tags that have as referent the real objects that we put at M1.

Having said that, it is easier to relate some of the BFO/OBI concepts with the XPDL classes to produce a mapping between elements of the two worlds. Table I presents the resulting mapping.

TABLE I
MAPPING BETWEEN XPDL AND BFO/OBI. THE RELEVANT CONCEPTS OF XPDL ARE MAPPED WITH CONCEPTS FROM BFO AND OBI.

Laboratory	XPDL	BFO	OBI
Protocol	Process	Directive information entity	Plan specification
Sub-protocol	SubFlow	Directive information entity	Plan specification
Unique single step of a protocol	Task/Tool	Directive information entity	Action specification
Real world (e.g., Illumina sample) or theoretical (e.g., Project) items	Data Type	independent continuant	material entity
		generically dependent continuant	information content entity
Objects properties	Data Field	specifically dependent continuant	quality

The main concept of *Protocol* is easily mapped to the workflow model by the notion of *Process*. In the XPDL specification a process is defined as a "combination of various activities with a specified flow of execution". An internal process consists of one or more activities, each comprising a logical, self-contained unit of work". We connected this concept with the OBI concept of *Plan specification*, defined as a "directive information entity that when construct it is realized in a process in which the bearer tries to achieve the objectives, in part by taking the specified actions. Plan speci-

 ISBN: 978-1-61208-181-6

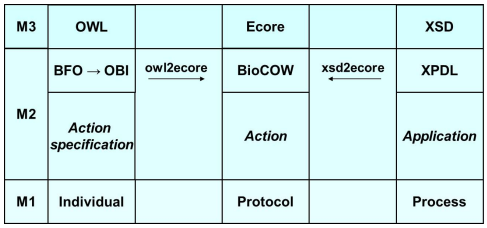


Fig. 3. The BioCOW meta-model is built combining XPDL with BFO/OBI. A standard XSD to Ecore transformation is used for XPDL. For BFO/OBI it has been used an existing tool dealing with OWL to Ecore transformation. As an example we show how the concepts of Action specification (XPDL) and Application (XPDL) are mapped into Action (BioCOW).

fications includes parts such as objective specification, action specifications and conditional specifications. A SubFlow (sub-protocol) is a process itself hence the mapping is the same as for process (i.e, *Plan Specification*).

The second main concept is the notion of unit of work. In XPDL this is backed by the Activity class, which can be of different kinds. One of those is the Task/Tool class, a service or an application required and invoked by the process. In the XPDL metamodel every tool declares a set of Applications. We mapped this XPDL concept with the Action specification in OBI, which defines it as a "directive information entity that describes an action the bearer will take".

Since an Activity is an atomic piece of work that may modify relevant data (declared as DataFields) we mapped on it both the XPDL concept of DataType and DataField. A Datatype in our model could be, aside from standard type, a *OBI:material entity* or an *OBI:information content entity*. We chose to map a DataField with the OBI concept of *quality*.

D. Implementation

To build the described meta-model we used the technology provided by the EMF and Ecore in particular. EMF provides tools to automatically convert heterogeneous formats to Ecore. Specifically there is a standard way to translate an XML Schema Definition (XSD) file in the Ecore format. Since XPDL is formulated in XSD we automatically imported it in EMF.

For the transformation of the OBI ontology into the Ecore format we followed the approach proposed by Hillairet et al. [10]. In particular, we translated the whole BFO ontology and the main classes of OBI from OWL. Using that approach we were able to manipulate both the ontology and the XPDL meta-model in a coherent way inside the EMF framework.

Our meta-model is built using as reference the XPDL meta-model. In order to actually concretize the mapping between XPDL and BFO/OBI we created a new class for every mapped classes. That new class inherits both the XPDL and BFO/OBI class as specified in the mapping shown in table I. For example, the *BioCOW:Action* class has, as a superclass, the *BFO:GenericallyDependentContinuant* class. It is worth noting that we have not specialized directly the XPDL meta-model since it is richer than we need for our purposes. We therefore based our model on XPDL retaining the main concepts and leaving out all the surplus details.

Using the resulting BioCOW (Bio-medicine Combined Ontology [and] Workflow) meta-model, we are now able to describe laboratory protocols in a formal yet intuitive way. By means of the Obeo designer we are able to build a Graphical User Interface (GUI) which associates graphicals symbols with constructs of the BioCOW meta-model. The efforts required to produce a GUI are greatly reduced using Obeo in contrast for example to the Eclipse Graphical Modeling Framework (GMF). Interestingly, however, Obeo still bases on GMF. The graphical editor enables the user to visually specify the desired protocols assembling components from the provided high-level language. In this manner, the designed protocols constructively conform to our meta-model.

E. Assessment

We are currently evaluating the BioCow meta-model in a real-world laboratory environment with the help of domain experts. We are comparing different frameworks analysing protocols widely used in the laboratory under the dimension of the *language* and *mediation* features.

IV. Discussion

The direction of our work relies on the potential of using ontology technologies in a MDE context. Thanks to the OTSs we can, for example, enable automatic reasoning for model consistency checking. Semantically assisted design (SAD) will allow the adoption of "intelligent" editors. Another important prospective is to enable systematic reuse of community-level shared formalized knowledge.

We place ourselves in the framework of the Features Model of Bridging MMTS and OTSs sketched in [12]. In our current work we have focused on some of those desiderata like the *mediation*. In fact, we have built a *mapping* from two specific modeling spaces. (XPDL and OBI) onto two technical spaces (MMTS and OTS). As part of that effort we were able to *integrate* concepts of XPDL with concepts of OBI. Working under the EMF toolbox we were able to incorporate some of the features of MMTS under a common technological and conceptual framework.

References

[1] J.-M. Favre and T. Nguyen, "Towards a megamodel to model software evolution through transformations," in *SETRA Workshop, Elsevier ENCTS*, vol. 127, 2004, pp. 59–74.

[2] E. Seidewitz, "What models mean," *Software, IEEE*, vol. 20, no. 5, pp. 26 – 32, sep. 2003.

[3] M. Pidd, *Tools for Thinking: Modelling in Management Science*, 3rd ed. Wiley, Feb. 2009. [Online]. Available: http://www.worldcat.org/isbn/0470721421

[4] U. Assmann, S. Zschaler, and G. Wagner, "Ontologies, Meta-models, and the Model-Driven Paradigm," *Ontologies for Software Engineering and Software Technology*, pp. 249–273, 2006.

[5] T. R. Gruber, "Towards principles for the design of ontologies used for knowledge sharing," in *Formal Ontology in Conceptual Analysis and Knowledge Representation*, N. Guarino and R. Poli, Eds. Deventer, The Netherlands: Kluwer Academic Publishers, 1993. [Online]. Available: citeseer.ist.psu.edu/gruber93toward.html

 ISBN: 978-1-61208-181-6

[6] Horrocks, I., Schneider, Patel P., and van Harmelen, F., "From shiq and RDF to OWL: The making of a web ontology language," *Journal of Web Semantics*, vol. 1, no. 1, pp. 7–26, 2003. [Online]. Available: {http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.2.7039}

[7] S. Schulz, M. Boeker, and H. Stenzhorn, "How granularity issues concern biomedical ontology integration," *Studies in health technology and informatics*, vol. 136, pp. 863–8, 2008.

[8] J. Bezivin, V. Devedzic, D. Djuric, J. Favreau, D. Gasevic, and F. Jouault, "An m3-neutral infrastructure for bridging model engineering and ontology engineering," in *Int. Conf. on Interoperability of Enterprise Software and Applications (INTEROP-ESA)*, Springer, Ed. Geneva, Switzerland: Springer, 2005, pp. 159–171, 1-84628-151-2. [Online]. Available: http://www.isima.fr/~favreau/files/publications/INTEROP-ESA.Extended.pdf

[9] E. Kapsammer, H. Kargl, G. Kramler, T. Reiter, W. Retschitzegger, and M. Wimmer, "Lifting metamodels to ontologies - a step to the semantic integration of modeling languages," in *In Proceedings of the ACM/IEEE 9th International Conference on Model Driven Engineering Languages and Systems (MoDELS/UML 2006*. Springer, 2006, pp. 528–542.

[10] B. F. Hillairet Guillaume and L. J. Yves, "Bridging emf applications and rdf data sources," in *Proceedings of the 4th international workshop on Semantic Web Enabled Software Engineering (SWESE) at ISWC'08*, 10 2008, pp. 26–40.

[11] F. S. Parreiras, S. Staab, and A. Winter, "On marrying ontological and metamodeling technical spaces," in *Proceedings of the the 6th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering*, ser. ESEC-FSE '07. New York, NY, USA: ACM, 2007, pp. 439–448. [Online]. Available: http://doi.acm.org/10.1145/1287624.1287687

[12] I. Kurtev, J. Bézivin, and M. Aksit, "Technological spaces: An initial appraisal," in *CoopIS, DOA 2002 Federated Conferences, Industrial track*, 2002. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.109.332

[13] L. Soldatova, W. Aubrey, R. King, and A. Clare, "The exact description of biomedical protocols," *Bioinformatics (Oxford, England)*, vol. 24, pp. i295–303, Jul 2008, 10.1093/bioinformatics/btn156.

[14] A. Maccagnan, M. Riva, E. Feltrin, B. Simionati, T. Vardanega, G. Valle, and N. Cannata, "Combining ontologies and workflows to design formal protocols for biological laboratories," *Automated Experimentation*, vol. 2, no. 1, p. 3, 2010.

[15] M. Courtot, W. Bug, F. Gibson, A. L. Lister, J. Malone, D. Schober, R. Brinkman, and A. Ruttenberg, "The owl of biomedical investigations," in *OWLED*, ser. CEUR Workshop Proceedings, C. Dolbear, A. Ruttenberg, and U. Sattler, Eds., vol. 432. CEUR-WS.org, 2008, p. xx.

[16] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. Goldberg, K. Eilbeck, A. Ireland, C. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. Scheuermann, N. Shah, P. Whetzel, and S. Lewis, "The obo foundry: coordinated evolution of ontologies to support biomedical data integration," *Nature Biotechnology*, vol. 25, pp. 1251–1255, Nov 2007, 10.1038/nbt1346.

[17] N. Russell, A. H. Hofstede, D. Edmond, and W. M. der Aalst, *Workflow Data Patterns: Identification, Representation and Tool Support*. Berlin/Heidelberg: Springer-Verlag, 2005, vol. 3716, ch. Chapter 23, pp. 353–368.

 ISBN: 978-1-61208-181-6

Modular Verification of Inter-enterprise Business Processes

Kais Klai
LIPN, CNRS UMR 7030
Université Paris 13
99 avenue Jean-Baptiste Clément
F-93430 Villetaneuse, France
Email: kais.klai@lipn.univ-paris13.fr

Hanen Ochi
LIPN, CNRS UMR 7030
Université Paris 13
99 avenue Jean-Baptiste Clément
F-93430 Villetaneuse, France
Email: hanen.ochi@lipn.univ-paris13.fr

***Abstract*— In this paper, we propose to adapt the Symbolic Observation Graphs (SOG) based approach in order to abstract, to compose and to check Inter-Enterprise Business Processes (IEBP). Each component (local process) is represented by a SOG where only the collaboration actions of the process are visible while its local behavior and its private structure are hidden. The entire IEBP is then abstracted by the composition of the components' abstractions (i.e., their SOGs). The main result of this paper is to demonstrate that the composition of the SOGs is deadlock free if and only if the original IEBP is deadlock free. We implemented our adaptation of the SOG construction and compared our abstraction and modular verification approach with the Operating Guidelines technique. The obtained results strengthen our belief that the SOGs are suitable to abstract and compose business processes especially when these are loosely coupled.**

Keywords-Process composition; abstraction; verification; deadlock-freeness.

I. INTRODUCTION

Recently, the trend in software architecture is to build Inter-Enterprise Business Processes (IEBP) modularly: Each process is designed separately and then the whole IEBP is obtained by composition. Even if such a modular approach is intuitive and facilitates the design problem, it poses two main problems: First, it is necessary to find an abstraction of the process that respect the privacy of the underlying enterprise (by hiring its internal organisation) and, at the same time, that supply enough information allowing to decide whether the collaboration with some partner is possible or not (safe or not). The second problem is that the correct behavior of each business process of the IEBP taken alone does not guarantee a correct behavior of the composed IEBP (i.e., properties are not preserved by composition). Thus, based on the abstraction of two (or more) processes, we shoud be able to say whether the composed process has the desired behavior or not (in our case, is deadlock free or not).

Proving correctness of the (unknown) composed process is strongly related to the model checking problem of a system model. Among others, the *symbolic observation graph* [5] based approach has proven to be very helpful for efficient model checking in general. Since it is heavily based on abstraction techniques and thus hides detailed information about system components that are not relevant for the correctness decision, it is promising to transfer this concept to the problem rised in this paper.

A SOG is a graph whose construction is guided by a subset of *observed* actions. The nodes of a SOG are *aggregates* hiding a set of states which are connected with non observed actions. The arcs of a SOG are exclusively labeled with observed actions.

The work presented in this paper is in line with those presented in [1] and [2]: How to adapt the SOG's structure in order to abstract and to compose business processes. Such an adaptation is achieved by attaching to each aggregate a (locally computed) sufficient and necessary information for detecting deadlocks that are possibly caused by the composition. The main contribution of this paper is to design a symbolic algorithm (based on sets operations) allowing an efficient computation of this information. This allowed to strengthen the conviction that SOGs represent a suitable abstraction since it respects the consraints mentioned above. Indeed, by observing only the collaborative activities of a process, publishing the corresponding SOG allows to hide its internal strcuture. The analysis power of the SOGs allows, in addition, to check the correctness of a composite process. The composition of SOGs is immediately suitable for synchronously composed processes. However, we can consider asynchronous composition as described in [1], [2]. The key idea is, when combining two processes, to involve a third process, representing the interface, into the composition stage. Such a component consists of buffers and the corresponding sending and receiving actions. Taken separately, this component is an infinite state system in the general case. However, since the whole IEBP is supposed to have finitely many states, only its reachable part is visited during an on-the-fly composition with both processes.

This paper is organized as follows: Section II presents some preliminary notions on WF-nets and labeled transition systems. Section III recalls the symbolic observation graphs and their application on business processes. Composition of SOGs and checking the deadlock freeness on the obtained synchronised product is the issue of Section IV. Section V is devoted to the implementation of our approach in addition to

 ISBN: 978-1-61208-181-6

experimental results. In Section VI, we discuss some related works and compare our technique to existing ones. Finally, Section VII concludes the paper and presents some aspects of the future work.

II. PRELIMINARIES

The technique presented in this paper applies to various kinds of process models that can map to labeled transition systems, e.g., Petri nets and, in particular, WorkFlow Petri nets (WF-nets) [11]. Since other modeling languages, which are more frequently used in practice, map to Petri nets, our approach is relevant for a very broad class of modeling languages. Applying our technique does not mean to construct labeled transition systems explicitly. Instead, abstractions of labeled transition systems are directly constructed from the original process models.

Definition 1 (Labeled Transition System):
A Labeled Transition System (LTS for short) is a $5-tuple < \Gamma, Act, \rightarrow, I, F >$ *where* :

- Γ *is a finite set of states*;
- Act *is a finite set of actions*;
- $\rightarrow \subseteq \Gamma \times Act \times \Gamma$ *is a transition relation*;
- $I \subseteq \Gamma$ *is a set of initial states*;
- $F \subseteq \Gamma$ *is a set of final states*.

In this paper, we restrain the set of states Γ to those that are reachable from the initial state. Moreover, we assume that a final state f is terminal (it has no successor) and that the set of actions Act is partionned into two disjoint subsets Obs (observed actions) and $UnObs$ (unobserved actions).
Below, we present some useful notations:

- For $s, s' \in \Gamma$ and $a \in Act$, we denote by $s \xrightarrow{a} s'$ that $(s,a,s') \in \rightarrow$.
- If $\sigma = a_1 a_2 \cdots a_n$ is a sequence of actions, $\overline{\sigma}$ denotes the set of actions occurring in σ, while $|\sigma|$ denotes the length of σ. $s \xrightarrow{\sigma} s'$ denotes that $\exists s_1, s_2, \cdots s_{n-1} \in \Gamma$: $s \xrightarrow{a_1} s_1 \xrightarrow{a_2} \cdots s_{n-1} \xrightarrow{a_n} s'$.
- The set *Enable(s)* denotes the set of actions a such that $s \xrightarrow{a} s'$ for some state s'. For a set of states S, *Enable(S)* denotes $\bigcup_{s \in S} Enable(s)$.
- $\pi = s_0 \xrightarrow{a_1} s_1 \xrightarrow{a_2} \cdots$ is used to denote a path of a LTS.
- $s \not\rightarrow$, for $s \in (\Gamma \setminus F)$, denotes that s is a dead state, i.e., $Enable(s) = \emptyset$.
- $Sat(s) = \{s' \mid s \xrightarrow{\sigma} s' \wedge \overline{\sigma} \subseteq UnObs\}$ is the set of states that are reachable from a state s by unobserved actions only. For $S \subseteq \Gamma$, $Sat(S) = \bigcup_{s \in S} Sat(s)$.
- $s \not\Rightarrow$, for $s \in \Gamma$, denotes that either no final state in F is reachable from s, or no state of $Sat(s)$ enables an observed action, i.e., $Enable(Sat(s)) \cap Obs = \emptyset$. Conversely, $s \Rightarrow$ denotes $\neg(s \not\Rightarrow)$.
- A finite path $C = s_1 \xrightarrow{\sigma} s_n$ is said to be a *cycle* if $s_n = s_1$ and $|\sigma| \geq 1$.
  If $\overline{\sigma} \subseteq UnObs$ then C is said to be a *livelock*. If, in addition, $s_1 \not\Rightarrow$ then C is called a *strong livelock* (a terminal cycle). Otherwise it is called a *weak livelock*.

If $s \not\Rightarrow$ for $s \in (\Gamma \setminus F)$, only a dead state or a *strong livelock* are reachable from s. In this paper we assume that a strong livelock behavior is equivalent to a deadlock. These two behaviors are not distinguished and both are called deadlock.

Consequently, if we want to check whether a given state s is a dead state or not, we need to check wether the predicate $s \not\Rightarrow$ holds or not. We say that we are interested in the *observed behavior* of s: (1) could s lead to the firing of some observed transitions in the future? (2) could s lead to a final state in the future? For this purpose, a *virtual* observed action, called *term*, is added to the observed actions Obs, it mentions that the system terminates proprely. The *Observed behavior*, namely λ, is then defined as a particular mapping applied to the set of states of a LTS as follows:

Definition 2 (Observed behavior mapping):
Let $\mathcal{T} = \langle \Gamma, Obs \cup UnObs, \rightarrow, I, F \rangle$ be a LTS. We define:

1) $\lambda_{\mathcal{T}} : \Gamma \rightarrow 2^{Obs}$
$$\lambda_{\mathcal{T}}(s) = \begin{cases} (Enable(Sat(s)) \cap Obs) \cup \{term\} \\ \text{if } F \cap Sat(s) \neq \emptyset \\ (Enable(Sat(s)) \cap Obs) \text{ otherwise} \end{cases}$$
2) $\lambda^{\cap}_{\mathcal{T}} : 2^{\Gamma} \rightarrow 2^{Obs}$
$\lambda^{\cap}_{\mathcal{T}}(S) = \bigcap_{s \in S} \lambda_{\mathcal{T}}(s)$
3) $\lambda^{\subseteq}_{\mathcal{T}} : 2^{\Gamma} \rightarrow 2^{2^{Obs}}$
$\lambda^{\subseteq}_{\mathcal{T}}(S) = \{\lambda^{\cap}_{\mathcal{T}}(Q) \mid \emptyset \subset Q \subseteq S\}$
4) $\lambda^{\min}_{\mathcal{T}} : 2^{\Gamma} \rightarrow 2^{2^{Obs}}$
$\lambda^{\min}_{\mathcal{T}}(S) = \{X \in \lambda^{\subseteq}_{\mathcal{T}}(S) \mid \nexists Y \in \lambda^{\subseteq}_{\mathcal{T}}(S) : Y \subset (X \setminus \{term\})\}$

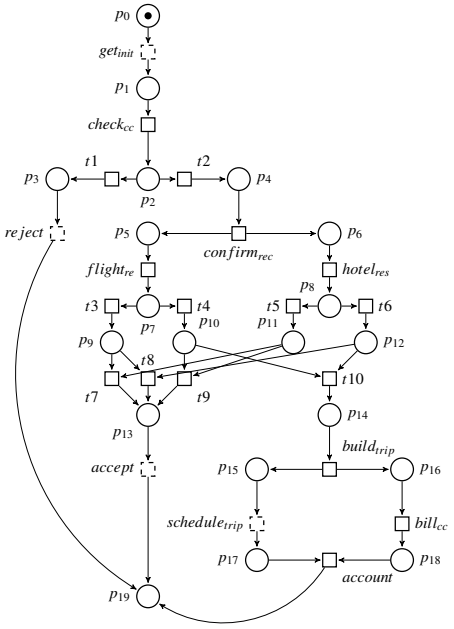
Informally, for each state s of a LTS $\mathcal{T}$, (1) the observed behavior of s, $\lambda_{\mathcal{T}}(s)$, stands for the set of observed actions which can be executed from s, possibly via a sequence of unobserved actions. In addition, *term* is a member of $\lambda_{\mathcal{T}}(s)$ if and only if a final state is reachable from s using unobserved actions only. (2) The observed behavior $\lambda^{\cap}_{\mathcal{T}}$ associated with a set of states S is the intersection of the observed behaviors of its elements. It contains the set of observed actions that are possible from each state of S. (3) $\lambda^{\subseteq}_{\mathcal{T}}(S)$ is a set of sets of observed actions such that each is the result of $\lambda^{\cap}_{\mathcal{T}}$ applied to a nonempty subset Q of S. (4) Finally, $\lambda^{\min}_{\mathcal{T}}(S)$ contains the minimal subsets of $\lambda^{\subseteq}_{\mathcal{T}}(S)$ w.r.t. the inclusion relation not concerning the *term* action. For instance, if there exist two states $s, s' \in S$ such that $\lambda(s) = \emptyset$ and $\lambda(s') = \{term\}$, then both subsets would appear as elements of $\lambda^{\min}_{\mathcal{T}}(S)$. This allows to distinguish whether a dead state or a final state is reached in S (in this case both kinds of state are reachable).

Thanks to the observed behavior, the deadlock freeness of a LTS can be reduced to check whether there exists a state s such that $\lambda_{\mathcal{T}}(s)$ contains the empty set. Similarly, a final state is reachable from a given state s iff *term* belongs to its observed behavior.

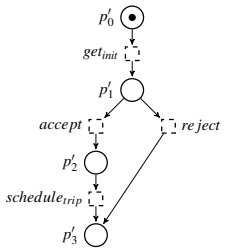
 ISBN: 978-1-61208-181-6

Running example :

We use an example of two business processes (taken from [9]), the trip reservation and the costumer, to illustrate the problem raised in this work. Figure 1(a) illustrates the WF-net associated with the trip reservation's process while Figure1(b) illustrates the WF-net associated with a costumer's process. The corresponding LTSs contain 13 nodes and 36 edges, and 4 nodes and 4 edges, respectively. We choosed such two examples in order to illustrate how SOG-based approach depends on the number of observed actions. In fact, the first contains a big proportion of unboserved actions (17/21), while, in the second all actions are observed. The two processes can collaborate (using dashed transitions) in order to form an IEBP.



(a) WF-net of trip reservation



(b) WF-net of customer

Figure 1. The WF-nets of a trip reservation and a costumer

III. SOG : SYMBOLIC OBSERVATION GRAPH

In this section, we recall the formal definition of a SOG associated with a LTS. We first define what is an *aggregate*: a node of the SOG. Compared to the first definition of SOGs (see [5]), the aggregates are here completed with the *observed behavior* of the hidden states. We will establish that this is the sufficient and necessary information allowing to detect possible deadlock states that can appear by composition. Recall that the deadlock freeness property is not preserved by composition: two deadlock free processes could lead, after composition, to a composite process with a dead state.

Definition 3 (aggregate):

Let $\mathcal{T} = \langle \Gamma, Act, \rightarrow, s_0, F \rangle$ be a labeled transition system with $Act = Obs \cup UnObs$. An *aggregate* is a couple $a = \langle S, \lambda \rangle$ defined as follows:

1) S is a nonempty subset of Γ s.t. $s \in S \Rightarrow Sat(s) \subseteq S$;
2) $\lambda = \lambda^{\min}_{\mathcal{T}}(S)$.

Informally, an aggregate a is defined as a couple (S, λ) where $a.S$ is its set of states (connected with unobserved actions) and $a.\lambda$ its observed behavior. The observed behavior associated with an aggregate a can help to know whether a contains a dead state ($\emptyset \in a.\lambda$) as well as whether a final state belongs to a ($\exists S' \subseteq a.\lambda$ s.t. $term \in S'$). In Section V, we propose a symbolic (set-based) algorithm allowing to efficiently compute the observed behavior of an aggregate.

Definition 4 (Symbolic Observation Graph):

A *symbolic observation graph SOG($\mathcal{T}$)* associated with a LTS $\mathcal{T} = \langle \Gamma, Obs \cup UnObs, \rightarrow, I, F \rangle$ is a LTS $\langle \mathcal{A}, Act', \rightarrow', I', F' \rangle$ such that:

1) $\mathcal{A}$ is a finite set of aggregates s.t.:
   a) There is an aggregate $a_0 \in \mathcal{A}$ s.t. $a_0.S = Sat(I)$;
   b) For each $a \in \mathcal{A}$ and for each $o \in Obs$ the set $\{s' \notin a.S \mid \exists s \in a.S, s \xrightarrow{o} s'\}$ is not empty if and only if it is a pairwise disjoint union of nonempty sets $S_1 \ldots S_k$ and for $i = 1 \ldots k$, there is an aggregate $a_i \in \mathcal{A}$ s.t. $a_i.S = Sat(S_i)$ and $(a, o, a_i) \in \rightarrow'$;
   c) For each aggregate $a \in \mathcal{A}$, the $a.\lambda$ attribute is computed following Definition 3;
2) $Act' = Obs$;
3) $\rightarrow' \subseteq \Gamma' \times Act' \times \Gamma'$ is the transition relation, obtained by applying 1b;
4) $I' = \{a_0\}$ (s.t. $a_0.S = Sat(\{I\})$;
5) $F' = \{a \in \Gamma' \mid \exists Q \in a.\lambda; term \in Q\}$.

Point 1b of Definition 4 deserves explanation: Given an aggregate a and an observed action o, the set of successors obtained by firing o from states of $a.S$ is partioned in disjoint subsets. For each of these subsets, there exists an aggregate in the SOG obtained by saturation on the states of the subset. For each such an aggregate a' there exists an arc from a to a' labeled with o. Thus, the SOG is non deterministic: an aggregate could have two successors by the same observed transition.

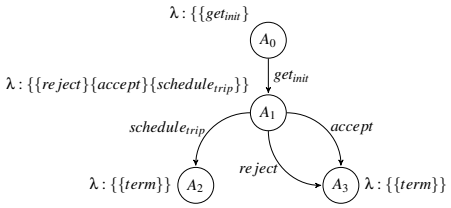
The construction of a SOG following the definition can be started by the initial aggregate a_0, then the SOG will be updated iteratively by adding new aggregates as long as the condition (1.b) is satisfied. Clearly, this construction is

not unique. One can take advantage of such a flexibility in order to obtain smaller aggregates (in terms of number of states). Even if the obtained SOG would have more aggregates in this case, it would consume less time and memory. This definition generalises the one given in [1], while the construction algorithm given in [5] is an example of implementation where the obtained graph is deterministic.

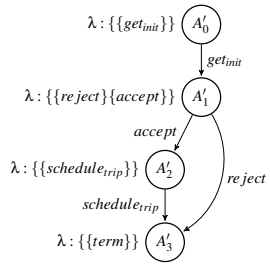
Notice that, once the SOG is built, the set of states of each aggregate has not to still be stored in memory any more. The unique useful information is the observed behavior annoting each node. A SOG is said to be deadlock free if none of its nodes admits the empty set as a member of its observed behavior.

The following proposition establishes that checking deadlock freeness of a SOG is equivalent to check deadlock freeness on the associated process (represented by its LTS)

Proposition 1: Let $\mathcal{W}$ be a business process, let $\mathcal{T} = \langle \Gamma, Act = Obs \cup UnObs, \rightarrow, I, F \rangle$ be the labeled transition system of $\mathcal{W}$ and let $\mathcal{G}$ be a SOG of $\mathcal{T}$. Then, $\mathcal{W}$ is deadlock free if and only if $\mathcal{G}$ is deadlock free.



(a) A SOG of trip reservation



(b) A SOG of customer

Figure 2. Two SOGs of the running example models

Figure 2 shows the two SOGs associated with the WF-nets of Figure 1. Figure 2(a) illustrates the SOG of the reservation trip model while Figure 2(b) shows the SOG of the customer model. We note that the two SOGs are deadlock-free: None of the aggregates of each SOG contains a deadlock state. We recall that the reachabilty graphs of the trip reservation and the customer models contain 13 nodes and 36 edges, and 4 nodes and 4 edges, respectively. It is clear, through this example, that bigger is the number of observed actions, smaller is the size of the obtained SOG. Especially, when all the actions of the service are observed, the SOG is isomorphic to the reachabilty graph.

IV. Composition of SOGs

In this section, we tackle the main idea of this paper: Given two (ore more) business processes (each ignoring internal details about the other), how to check that their composition is deadlock free? We showed in the previous section that the SOG can represent a good abstraction of business processes on which the deadlock freeness property can be checked. Here, we prove that they can also be used in a compositional way: The composition of two SOGs can be useful to check the correctness of the composition of the underlying processes.

We propose to build the synchronized product of two (or more) SOGs so that the obtained graph remains a SOG. Now, the difficulty is to compute the observed behavior of the synchronized product SOG. In fact, the states abstracted by an aggregate are hidden (actually, they do not exist in memory any more, once the SOG is built) and directly computing their observed behavior is not possible. Thus, given two aggregates a_1 and a_2 belonging to two different SOGs, we propose to deduce the observed behavior of the product aggregate, $a = a_1 \times a_2$, from those of a_1 and a_2.

Definition 5 (aggregate product):
Let $\mathcal{T}_i = \langle \Gamma_i, Obs_i \cup UnObs_i, \rightarrow_i, I_i, F_i \rangle, i = 1,2$ be two LTSs. Let $a_i = \langle S_i, \lambda_i \rangle$ be two aggregates of two associated SOGs. The *product aggregate* $a = \langle S, \lambda \rangle$, denoted by $= a_1 \times a_2$, is defined by:

- $a.S = a_1.S \times a_2.S$;
- $a.\lambda = \{(x \cap y) \cup (x \cap (Obs_1 \setminus Obs_2)) \cup (y \cap (Obs_2 \setminus Obs_1)) \mid x \in a_1.\lambda,\ y \in a_2.\lambda\}$.

Notice that the *term* action is supposed to be shared by both LTSs. Intuitively, an observed action is possible from a state $s = (s_1, s_2)$ in $a = a_1 \times a_2$ if it is observed in $\mathcal{T}_1$ and $\mathcal{T}_2$ and possible from both states s_1 and s_2, or it is observed only in $\mathcal{T}_1$ (resp. $\mathcal{T}_2$) and possible from s_1 (resp. s_2).

Definition 6 (SOG synchronized product):
Let $\mathcal{T}_i = \langle \Gamma_i, Obs_i, \rightarrow_i, I_i, F_i \rangle, i = 1,2$ be two SOGs. The *synchronized product* of $\mathcal{T}_1$ and $\mathcal{T}_2$, denoted by $\mathcal{T}_1 \times \mathcal{T}_2$ is the SOG $\langle \Gamma, Obs, \rightarrow, I, F \rangle$ where:

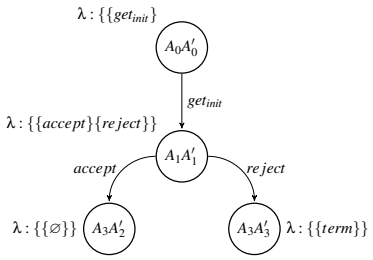
1) $\Gamma = \Gamma_1 \times \Gamma_2$;
2) $Obs = Obs_1 \cup Obs_2$;
3) $\rightarrow$ is the transition relation, defined by:
$$\forall (a_1, a_2) \in \Gamma' : (a_1, a_2) \xrightarrow{o} (a'_1, a'_2) \Leftrightarrow \begin{cases} a_1 \xrightarrow{o}_1 a'_1 \wedge a_2 \xrightarrow{o}_2 a'_2 & \text{if } o \in Obs_1 \cap Obs_2 \\ a_1 \xrightarrow{o}_1 a'_1 \wedge a_2 = a'_2 & \text{if } o \in Obs_1 \setminus Obs_2 \\ a_1 = a'_1 \wedge a_2 \xrightarrow{o}_2 a'_2 & \text{if } o \in Obs_2 \setminus Obs_1 \end{cases}$$
4) $I = I_1 \times I_2$;
5) $F = F_1 \times F_2$.

Note: The set of aggregates Γ' is reduced to the states that are reachable from the initial aggregate.

The following proposition establishes that the syncronized product of two SOGs is a SOG. This result combined with Proposition 1 allows to reduce the deadlock freeness verification of an IEBP to the verification of the synchronized

product of the SOGs derived from its components.

Proposition 2: Let $\mathcal{W}_i$, for $i \in \{1,2\}$, be two business processes, whose IEBP is $\mathcal{W}$, and let $\mathcal{T}_i$ be their corresponding LTSs. Let $\mathcal{G}_i$ be a SOG associated with T_i with respect to the set of observed actions Obs_i, and let $\mathcal{G}$ be the synchronized product of $\mathcal{G}_i$. Then $\mathcal{G}$ is a SOG of the $\mathcal{W}$'s LTS with respect to $Obs_1 \cup Obs_2$.



(a) the SOG synchronized product

Figure 3. the SOG synchronized product

Figure 3 illustrates the SOG obtained by synchronizing the SOGs of Figure 2. We note that it contains a deadlock aggregate $A_3A'_2$ although A_3 and A'_2 are deadlock-free. In fact, $\{term\} \cap \{schedule_{trip}\} = \emptyset$.

V. Implementation and Experimental Result

A. Implementation

The construction of the original version of SOG has been already implemented in [5] and a model checker of Linear Temporal Logic formulae based on SOGs was proposed in [6]. In this work, we adapted the existing tool to the context of composition of business processes. Thus, the main task was to adapt the construction of the SOG so that the observed behavior [2] is computed for each new aggregate. A direct implementation of the observed behavior of a given aggregate (following Definition 3) implies to consider each state belonging to the aggregate separately. This would considerably decrease the efficiency of the approach. In fact, each aggregate is encoded with a BDD and all the operations manipulating the aggregates should be based on set operations. Therefore, we have implemented an algorithm (see Algorithm 1) for the computation of the observed behavior that is exclusively based on set operations applied to the states of a given aggregate.

The input of Algorithm 1 are an aggregate A, the set of observed transitions Obs, the set of unobserved transitions $UnObs$ and the final set of states F. It computes the observed behavior associated with the aggregate A (i.e., $A.\lambda$).

We use a map (called R) whose elements are couples of sets of events and sets of states (line 1). Each element (O,S) satisfies the following: each state of S enables each transition of O. This map is progressively updated so that, at the end of the algorithm, the set of its keys (the first element of the couples) form the observed behavior of the aggregagte A (line 18). The first step of the algorithm (lines $2-4$) consists in: (1) checking whether a final state belongs to $A.S$, (2) if it is the case creating a new couple $(\{term\},S)$ where S is the set of the immediate predecessors of the final states present in A. The latter task is performed by using the $PreIm()$ function. The second step of the algorithm (lines $5-9$) allows to fill the map R with couples of the form $(\{o\},S)$ where o is an observed action and S the subset of states of A enabeling o. Once the map R is filled, it is analysed in the third part of the algorithm (lines $10-17$). The idea is to look between elements of R those having the same enabling sets of states (the second component of each couple). For each pair (O,S) and (O',S) in R the first couple is updated by adding O' to O while the second is removed from the map. Indeed, states in S enable each action in O or in O' and should be associated with the set $O \cup O'$.

The final part of the algorithm (lines $19-29$) is dedicated

Algorithm 1 Computing the Observed Behavior

Require: $AgregateA, Obs, UnObs, Setofstates F$
Ensure: $A.\lambda$

```
1:  Map < Set of events, Set of states > R
2:  if F ∩ A.S ≠ ∅ then
3:     insert ({term}, Pred(F, A.S, UnObs)) in R
4:  end if
5:  for o ∈ Obs do
6:     if Enable(A.S, o) ≠ ∅ then
7:        insert ({o}, Enable(A.S, o)) in R
8:     end if
9:  end for
10: for (O, S) ∈ R do
11:    for (O', S') ∈ R do
12:       if S = S' then
13:          (O, S) ← (O ∪ O', S)
14:          remove (O', S') from R
15:       end if
16:    end for
17: end for
18: λ ← Set of keys of R
19: Set of states E ← ∅
20: for t ∈ (Obs ∪ UnObs) do
21:    E ← E ∪ Enable(S, t)
22: end for
23: if E ≠ S then
24:    λ ← λ ∪ {∅}
25: else
26:    if (PreIm*(Enable(A.S, Obs) ∪ (F ∩ A.S), UnObs) ≠
       A.S) then
27:       λ ← λ ∪ {∅}
28:    end if
29: end if
30: return λ
```

Model	Places	Trans	Obs	RG		OG			SOG		
				States	Edges	States	Edges	time(s)	States	Edges	time(s)
C	18	11	4	26	66	12	20	<1	5	4	<1
SC	15	9	4	11	11	9	11	<1	7	7	<1
OS	15	8	8	10	10	12	17	<1	10	10	<1
R	38	33	17	28	33	369	14 e^2	<1	17	17	<1
Ph5	36	16	10	417	10 e^2	14 e^2	34 e^2	16	297	721	8
Ph6	43	19	12	14 e^2	46 e^2	61 e^2	17 e^3	245	991	28 e^2	42
Ph7	50	22	14	52 e^2	19 e^3	26 e^2	88 e^3	42 e^2	33 e^2	11 e^3	162
Ph10	71	31	20	23 e^5	23 e^4	-	-	-	12 e^4	58 e^4	15 e^2
Ph10	71	31	4	23 e^5	23 e^4	-	-	-	21	50	15

Table I
EXPERIMENTAL RESULTS: OG VS. SOG

to the analysis of the deadlock states inside the aggregate A. If a deadlock state is found in $A.S$ then the empty set is added to λ. A terminal state is detected (lines $19-24$) when the set of states enabling some transition (observed or not) is not equal to the whole set $A.S$. In order to detect strong livelocks (terminal cycles), we iterate on the $PreIm()$ function in order to compute all the states in $A.S$ that possibly lead either to a state in $Enable(A.S, Obs)$ (i.e., a state enabling some observed action), or to a final state. If the result is not equal to $A.S$ then there is a terminal cycle in A and the empty set should belong to $A.\lambda$ (line $26-27$).

In addition to the implementation of the observed behaviour algorithm, we integrated new functionnalities to allow the abstraction and the composition of business processes: Given a WF-net description of one or more business processes, it is possible to check the deadlock freeness property on the fly by building the correspondig SOG. The user can choose to stop the construction of the SOG as soon as a deadlock state is reached, or not. In the last case a textual description of the whole SOG is supplied.

B. Experimental results

We used our implementation in order to build the SOG associated with several business processes from different domains. We do not describe these models here because of lack of place but we give their WF-net models' size (in terms of number of places and transitions) as well as the size of their reachability state graphe RG (in terms of number of nodes and arcs) in Table I. These models were also supplied to Wendy ([10]), a tool to analyse interacting open nets. One of the fonctionalities of Wendy is to build the *Operating Guideline* [3], an annotated automata, in order to abstract a model and to check compatibility between two models (i.e., whether two models can collaborate safely). The corresponding results are illustrate in Table I (column OG)

The obtained results show clearly that SOGs-based approach outperforms than the operating guidelines-based one. The SOG is always (at least for the tested examples) smaller than the operating guideline graph and its construction faster. It is interesting to notice that the size of the operating guideline can be greater than the size of the reachability graph. For instance, this is the case of the online shop model (OS). The corresponding SOG is isomorphic to the reachability graph since all the transitions are observed. The SOG-based approach is especially efficient for loosely coupled models (with a few number of observed actions). This can be easily noticed if we look to the two last lines of Table I: at line 8 the 10 philosophers model is obtained by composition of 10 models (each represents one philosopher). Each philosopher contains two observed transitions (those allowing to pick up the forks) and the total number of observed transitions is 20 over 31. In the last line, however, this model is obtained by composing only two models, each representing five philosophers and contains 2 observed transitions (thus 4 observed transitions over 31). In this case the size of the SOG is negligible comparing to the first case. In both cases Wendy is not able to supply the result because of the explosion of the corresponding state space.

VI. RELATED WORK

The importance of dealing with business processes on one hand and business process composition on the other hand is reflected in the literature by several publications. Below, we discuss some related approaches.

The public-to-private approach introduced by W. van der Aalst in [12] consists of three steps. Firstly, the organizations involved agree on a common and sound public workflow, which serves as a contract between these organizations. Secondly, each task of the public workflow is mapped onto one of the domains (*i.e.*, organization). Each domain is responsible for a part of the public workflow, referred to its public part. Thirdly, each domain can now make use of its autonomy to create a private workflow. To satisfy the correctness of the overall inter-organizational workflow, however, each domain may only choose a private workflow which is a subclass of its public part [13]. The public-to-private approach allows to the local processes to be decoupled as much as possible and to have some degree of understanding about the nature of the interaction between the processes of

 ISBN: 978-1-61208-181-6

the different business partners. A problem to be encountered by this approach is confidentiality that prevents a complete view of local workflow. Indeed, to check the deadlock property, one needs the model of the global workflow. This model however is often not available for inter-organizational workflows since organizations are not willing to disclose their workflows (e.g., for privacy reasons). Therefore, our technique that abstracts local workflows using SOGs is well suited to verify properties and preserve organization privacy.

In [7], a formal model for services called service automota is defined by P. Massuthe and K. Schmidt. Based on this representation, the authors combine all the well-interactions between a service and its determinitic partners on an annotated automaton called Operating Guideline. This automaton characterizes all services wich interact properly with the corresponding service. This approach of abstraction was extended to composition of web services by P. Massuthe and K. Wolf in [8], allowing publishers on the web to maintain privacy of the services and to present only the essential behavior information for matching. Authors give a matching algorithm that can be applied between an operating guideline and a web service model and check whether the matching is possible or not.

Another approach for workflow matchmaking was proposed by A. Martens in [9]. It assumes that two workflows match if they are equivalent. To reach this end, the author introduces the notion of communication graph *c-graph* and usability graph (*u-graph*). If the *u-graph* of a workflow is isomorphic to the *c-graph* of another workflow, then the two workflows are considered equivalent.

In conclusion, to the best of our knowledge, none of the existing approaches combine symbolic (using BDDs) abstraction and modular verification to check the correctness of inter-organizational processes. They always deal with an explicite representation of the system's behavior, which accentuate the state space explosion problem.

VII. Conclusion

The main emphasis of this paper is on composition of business processes using symbolic observation graphs: How can we compose extended SOGs such that the resulting SOG is still small but represents the behavior of the IEBP in an appropriate way? We addressed the problem of checking correctness of IEBPs compositionally. We established that and how symbolic observation graphs can be extended and efficiently used for that purpose. We implemented the presented approach and compared the obtained results against the operating guidelines approach. The obtained results confirms our belief that the SOG is a suitable abstraction of business processes that offers, in addition, interesting analysis capabilities.

Our future work will be on studying other correction criteria (e.g., soundness) by depicting the necessar local information (like we did wit the observed behavior) to be stored (within each aggregate) so that the desired property can be checked on the composition of the obtained SOGs. We also plan to extend our approach to deal with resources.

References

[1] K. Klai, S. Tata and J. Desel *Symbolic Abstraction and Deadlock Freeness Verification of Inter-Enterprise Processes*, In Proceedings of the 29th International Conference On Business Process Management, Ulm, Germany, September 2009, 294–309, Springer-Verlag.

[2] K. Klai, S. Tata and J. Desel *Symbolic Abstraction and Deadlock-Freeness Verification of Inter-Enterprise Processes*, In Journal of Data & Knowledge Engeneering (DKE), 2011.

[3] N. Lohmann, P. Massuthe and K. Wolf *Operating Guidelines for Finite-State Services*, In Proceedings of Petri nets'07, Siedlce, Poland, June 25-29, 2007, 321-341, Springer-Verlag, Berlin, Heidelberg

[4] J. Koehlerand B. Srivastava *Web Service Composition: Current Solutions and Open Problems*, ICAPS 2003 Workshop on Planning for Web Services, pages 28 - 35.

[5] S. Haddad, JM. Ilié and K. Klai *Design and Evaluation of a Symbolic and Abstraction-based Model Checker*, In Proceedings of Automated Technology for Verification and Analysis: Second International Conference, ATVA 2004, Taipei, Taiwan, October 31-November 3, 2004,198–210, Springer LNCS 3299

[6] K. Klai and D. Poitrenaud *MC-SOG: An LTL Model Checker Based on Symbolic Observation Graphs*, In Proceedings of Petri Nets'08, Xian, China 2008, 288–306, Springer LNCS, 5062

[7] P. Massuthe and K. Schmidt *Operating Guidelines for Services*, In Proceedings of 12. Workshop "Algorithmen und Werkzeuge fr Petrinetze" September, 2005, 78–83

[8] P. Massuthe and K. Wolf *An Algorithm for Matching Nondeterministic Services with Operating Guidelines*, International Journal of Business Process Integration and Management 2007, Vol. 2, 81 - 90

[9] A. Martens *Usability of web services*, In Proceedings of the Fourth international conference on Web information systems engineering workshops, 2003, Roma, Italy, 182–190, IEEE Computer Society

[10] N. Lohmann and D. Weinberg *Wendy: A Tool to Synthesize Partners for Services*, In Proceedings of Petri Nets' 10, Braga, Portugal, June, 2010, 297-307

[11] W. Van der Aalst *The Application of Petri Nets to Workflow Management*, In Journal of Circuits, Systems, and Computers 1998, 21-66

[12] W. van der Aalst *Loosely Coupled Interorganizational Workflows: Modeling and Analyzing Workflows Crossing Organizational Boundaries*, In Journal of Information and Management, 2000, 67-75

[13] W. van der Aalst and M. Weske,*The P2P Approach to Interorganizational Workflows*, In Proceedings of the 13th International Conference on Advanced Information Systems Engineering, 2001, 140-156

 ISBN: 978-1-61208-181-6