LNBIP 7

Witold Abramowicz
Dieter Fensel (Eds.)

Business Information Systems

11th International Conference, BIS 2008
Innsbruck, Austria, May 2008
Proceedings

Springer

Lecture Notes
in Business Information Processing 7

Series Editors

Witold Abramowicz Dieter Fensel (Eds.)

Business Information Systems

11th International Conference, BIS 2008
Innsbruck, Austria, May 5-7, 2008
Proceedings

Springer

Volume Editors

Witold Abramowicz
Poznań University of Economics
Department of Information Systems
Al. Niepodległości 10, 60-967 Poznań, Poland
E-mail: W.Abramowicz@kie.ae.poznan.pl

Dieter Fensel
University of Innsbruck
STI Innsbruck
Technikerstrasse 21 a, 6020 Innsbruck, Austria
E-mail: dieter.fensel@sti2.at

Library of Congress Control Number: 2008925412

ACM Computing Classification (1998): J.1, H.4, D.2.12, H.5.3

ISSN 1865-1348
ISBN-10 3-540-79395-X Springer Berlin Heidelberg New York
ISBN-13 978-3-540-79395-3 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springer.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12256518 06/3180 5 4 3 2 1 0

Preface

BIS 2008 was the 11th in a series of international conferences on Business Information Systems. The conference took place in Innsbruck which means that after Klagenfurt it was the second Austrian BIS edition. The BIS conference series from its very roots has been recognized by professionals as a forum for the exchange and dissemination of topical research in the development, implementation, application and improvement of computer systems for business processes.

The theme of this conference was "Business Processes and Social Contexts—Reaching Beyond the Enterprise." The material collected in this volume covers research trends as well as current achievements and cutting-edge developments in the area of modern business information systems. A set of 41 papers were selected for the presentation during the main event and grouped around conference topics: Business Process Management, Service Discovery and Composition, Ontologies, Information Retrieval, Interoperability, Mobility and Contexts, Enterprise Resource Planning, Wikis and Folksonomies, Rules and Semantic Queries.

The Program Committee consisted of more than 80 members that carefully evaluated all the submitted papers. This year they were supported by an EasyChair review system, and again, we observed an increase in the quality of the reviews. This not only raised the quality of the conference but also positively affected the work of the authors.

The regular program was complemented by the outstanding keynote speakers. We are proud that BIS 2008 hosted Alistair Barros (SAP, Australia), Hans Ulrich Buhl (University of Augsburg, Germany), Fabio Ciravegna (University of Sheffield, UK), John Davies (BT, UK), and Frank Leymann (University of Stuttgart, Germany).

This year the special SUPER PhD Programme was a part of BIS. The event was organized in co-operation with Poznań University of Economics. The goal of the SUPER PhD Programme session was to create an opportunity for doctoral students to test their research ideas, present their current progress and future plans, and above all, to receive constructive criticism and insights related to their future work and research career perspectives.

BIS 2008 was kindly supported by SUPER Integrated Project (FP6-026850), Service Web 3.0 (FP7-216937) and Semantic Technology Institute International (Austria).

May 2008

Witold Abramowicz
Dieter Fensel

Conference Organization

Program Chairs

General Co-chairs	Witold Abramowicz, Poznań University of Economics, Poland
	Dieter Fensel, University of Innsbruck, Austria

Program Committee

Andreas Abecker	FZI, Germany
Ajith Abraham	Chung-Ang University, Korea
Dimitris Apostolou	University of Piraeus, Greece
Kalina Bontcheva	University of Sheffield, UK
Laszlo Böszörmenyi	University of Klagenfurt, Austria
Wojciech Cellary	Poznań University of Economics, Poland
Flavio Corradini	University of Camerino, Italy
Tommaso Di Noia	Technical University of Bari, Italy
Ying Ding	University of Innsbruck, Austria
Martin Dzbor	The Open University, UK
Bogdan Franczyk	University of Leipzig, Germany
Ulrich Frank	University of Duisburg-Essen, Germany
Johann-Christoph Freytag	Humboldt University, Germany
Fausto Giunchiglia	University of Trento, Italy
Jerzy Gołuchowski	University of Economics in Katowice, Poland
Jorge Marx Gómez	University of Oldenburg, Germany
Norbert Gronau	University of Potsdam, Germany
Oliver Günther	Humboldt University, Germany
Hele-Mai Haav	Tallinn University of Technology, Estonia
Manfred Hauswirth	National University of Ireland, Ireland
Armin Heinzl	University of Mannheim, Germany
Martin Hepp	University of Innsbruck, Austria
Knut Hinkelmann	University of Applied Sciences Northwestern Switzerland, Switzerland
Tomasz Kaczmarek	Poznań University of Economics, Poland
Paweł J. Kalczyński	CSU Fullerton, USA
Uzay Kaymak	Erasmus University Rotterdam, The Netherlands
Jerzy Kisielnicki	University of Warsaw, Poland
Gary Klein	University of Colorado at Colorado Springs, USA
Ralf Klischewski	German University in Cairo, Egypt
Jacek Koronacki	Polish Academy of Sciences, Poland

Ryszard Kowalczyk	Swinburne University of Technology, Australia
Marek Kowalkiewicz	SAP Research Brisbane, Australia
Helmut Krcmar	TU Müchen, Germany
Sanjay Kumar	XLRI School of Management, India
Dennis Kundisch	Albert-Ludwig University of Freiburg, Germany
Dominik Kuropka	University of Potsdam, Germany
Henry Linger	Monash University, Australia
Peter C. Lockemann	University of Karlsruhe, Germany
Peter Loos	University of Saarbrücken, Germany
Leszek Maciaszek	Macquarie University, Australia
Yannis Manolopoulos	Aristotle University, Greece
Heinrich C. Mayr	University of Klagenfurt, Austria
Marie-Francine Moens	Katholieke Universiteit Leuven, Belgium
Günter Müller	University of Freiburg, Germany
Ludwig Nastansky	University of Paderborn, Germany
Markus Nüttgens	University of Hamburg, Germany
Andreas Oberweis	University of Karlsruhe, Germany
Mitsunori Ogihara	University of Rochester, USA
Maria E. Orłowska	Polish-Japanese Institute of Information Technology, Poland
Jeff Pan	University of Aberdeen, UK
Marcin Paprzycki	Polish Academy of Sciences, Poland
Eric Paquet	National Research Council, Canada
Arnold Picot	University of Munich, Germany
Jakub Piskorski	EU, JRC, Web and Language Technology, Italy
Jaroslav Pokorný	Charles University, Czech Republic
Elke Pulvermüller	University of Osnabrück, Germany
Ulrich Reimer	University of Konstanz, Germany
Narcyz Roztocki	State University of New York at New Paltz, USA
Shazia Sadiq	The University of Queensland, Australia
Kurt Sandkuhl	Jönköping University, Sweden
Alexander Schill	Dresden University of Technology, Germany
Elmar J. Sinz	University of Bamberg, Germany
Janice C. Sipior	Villanova University, USA
Kilian Stoffel	University of Neuchâtel, Switzerland
York Sure	University of Karlsruhe, Germany
Witold Suryn	École de Technologie Supérieure, Canada
Vojtěch Svátek	VSE Praha, Czech Republic
Barbara Thönssen	University of Applied Sciences Northwestern Switzerland, Switzerland
Robert Tolksdorf	Free University Berlin, Germany
Susan Urban	Texas Tech University, USA
Olegas Vasilecas	Vilnius Gediminas Technical University, Lithuania

Herna Viktor	University of Ottawa, Canada
Raphael Volz	FZI, Germany
Hannes Werthner	Vienna University of Technology, Austria
Mathias Weske	Hasso Plattner Institute for It-Systems Engineering, Germany
Krzysztof Węcel	Poznań University of Economics, Poland
Viacheslav E. Wolfengagen	Institute JurInfoR-MSU, Russia
Stanisław Wrycza	University of Gdansk, Poland
Vilas Wuwongse	Asian Institute of Technology, Thailand
Adnan Yazici	Middle East Technical University, Turkey
Sławomir Zadrożny	Polish Academy of Sciences, Poland
Jozef Zurada	University of Louisville, USA

External Reviewers

Mohan Baruwal Chhetri	Sami Bhiri	Veli Bicer
Jinjun Chen	Tobias Conte	Francesco De Angelis
Peter Fettke	Walid Gaaloul	Christoph Goebel
Christoph Gruen	David Heise	Hanna Krasnova
Philipp Langer	Algirdas Laukaitis	Philipp Liegl
Mathias Lux	Daniele Maggiore	Apostolos Papadopoulos
Azzurra Ragone	Barbara Re	Christoph Riedl
Oliviero Riganelli	Michele Ruta	Rainer Schuster
Armin Sharafi	Sergejus Sosunovas	Jonas Sprenger
Stefan Strecker	Stuart Taylor	Eufemia Tinelli
Christoph Tribowski	Denny Vrandecic	Valentin Zacharias
Marco Zapletal		

Organizers

Poznań University of Economics
Department of Information Systems
Al. Niepodległości 10
60-967 Poznań
Poland
phone: +48(61)854-3381
fax: +48(61)854-3633
http://www.kie.ae.poznan.pl

University of Innsbruck
STI Innsbruck
Technikerstrasse 21a
6020 Innsbruck
Austria
phone: +43(512)507-6488
fax: +43(512)507-9872
http://www.sti-innsbruck.at

Organizing Committee

Dominik Flejter
Konstanty Haniewicz
Piotr Stolarski
Krzysztof Węcel
Ilona Zaremba

Table of Contents

Information Retrieval

Wikis and Folksonomies

Ontology Development

Ontology Application

BPM: Modelling

BPM: Towards Flexibility

BPM: Monitoring and Verification

BPM: Involving People

Service Discovery and Composition

Interoperability

ERP: Analyses

ERP: International Experiences

Rules and Semantic Queries

Mobility and Contexts

An Entry Vocabulary Module for a Political Science Test Collection

Benjamin Berghaus[1], Thomas Mandl[1], Christa Womser-Hacker[1], and Michael Kluck[2]

[1] Information Science, University of Hildesheim
Marienburger Platz 22, D-31141 Hildesheim, Germany
mandl@uni-hildesheim.de
[2] Stiftung Wissenschaft und Politik, SWP
Ludwigkirchplatz 3-4, 10719 Berlin, Germany

Abstract. We describe the design of a retrieval test for texts on political science. The corpus contains 600,000 documents in various languages. A set of 25 typical topics for the domain was developed and relevance judgments were provided by domain experts. To improve the domain specific retrieval performance, an entry vocabulary module (EVM) which maps query terms to the domain specific vocabulary was developed. We compare a base run to a blind relevance, feedback run as well as to both a static and a dynamic EVM. The dynamic EVM is presented in this paper. It can be shown that the dynamic EVM greatly improves recall and also improves precision. An innovative topic specific analysis proves that the EVM also hurts some topics.

Keywords: Domain specific, evaluation issues, entry vocabulary, test collections, query analysis.

1 Introduction

The importance of testing retrieval system in environment of use has often been stressed. Domain specific terminology poses many challenges to systems. The semantic heterogeneity between full text terms, user query terms and dedicated thesauri leads to mismatches between queries and documents. The disparity of ontologies has often been discussed as a major problem for mutual understanding and the sharing of knowledge. For users of retrieval systems, it is necessary that the systems understand their language and that they can access the information needed with their own terminology. Users should not be required to learn the language of the documents for querying a system.

Most knowledge domains have been organized into ontologies several times form different perspectives. Each perspective is justified within its own context. Information system engineers are faced with a large variety of ontologies for each domain. Sharing knowledge across different perspectives represented by semantically heterogeneous ontologies remains tedious. This challenge of semantic heterogeneity has led to many intellectual and technological solutions for different forms of semantic unification [6].

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 1–11, 2008.

Heterogeneous ontologies occur in many areas. The most typical attempt to resolve this problem is standardization and the concentration on one ontology only. However, this may not always be possible and aspects of the domain may get lost. Some of the problems arising from ontology mismatch are presented here with examples from information science:

- Challenges due to different terminology (e.g. usability, human-computer interaction, or interface design)
- Different hierarchy formation or poly-hierarchical structures (e.g. information science -> information retrieval -> information retrieval evaluation or information science -> evaluation -> information retrieval evaluation)
- Different assignment of segments (e.g. information retrieval -> user interfaces for information retrieval or user interfaces -> user interfaces for information retrieval)

The remainder of this paper is organized as follows. The next section provides a brief overview of the research on domain specific evaluation results and their validity. Section three describes the data for the test design developed for this study. In section four, the entry vocabulary module is discussed. Subsequently, results are presented and discussed.

2 Domain Specific Retrieval Evaluation

The evaluation of domain specific retrieval systems has been promoted by all three major evaluation campaigns. The Text Retrieval Conference (TREC) has initiated a track on genetic information, the Asian initiative NTCIR has evaluated technical patents and the Cross Language Evaluation Forum (CLEF) [10] has introduced social science data [8].

The validity of large-scale information retrieval experiments has been the subject of a considerable amount of research. Zobel concluded that the TREC (Text REtrieval Conference) experiments are reliable as far as the ranking of the systems is concerned [17]. Buckley & Voorhees have analyzed the reliability of experiments for different sizes of the topic set [3]. They concluded that the typical size of the topic set of some 50 topics in TREC is sufficient for high test reliability. Several different methods are applied for the reliability analysis. Traditional statistical significance testing has a long tradition in information retrieval. However, more recent analysis is oriented toward the use of the data. Retrieval evaluation results are needed to distinguish between two systems. As a consequence, the so called swap rate is applied. It calculates how many positions in the systems ranking change when a different or smaller topic set is used for the test and for ranking the systems. These results have hinted that the statistical tests overestimate the potential error and that often smaller differences between systems should be considered as significant. Even smaller test set usually lead to a satisfactory level of reliability. We can assume that the test described in this paper also leads to fairly reliable results in determining which algorithms produce good retrieval results.

3 Evaluation Design

The design of the retrieval test presented here follows the model of the TREC and CLEF [10]. A document collection, a set of topics representing natural user information needs and subsequently relevance assessment by humans for a data set. Contrary to the typical evaluation campaigns, domain specific data requires that domain experts judge the relevance of the documents. This has been the case in the study presented here.

The data for this retrieval test was provided by the *Stiftung Wissenschaft und Politik* (SWP), Berlin, a German think tank which supports the German Parliament, the Ministry of Foreign Affairs, and further ministries. For this purpose, SWP is continuously building a high quality database of documents on international relations and security affairs.

The collection comprises 600,000 documents on topics such as international policy, international economy and security issues. Half of the documents deal with European countries and institutions. All other world regions are also represented. One fourth of the documents contain full text, others consist of title and abstracts. All documents have been intellectually indexed by information specialists who assigned terms from a domain specific thesaurus [4]. The terms assigned fall into different categories. The content is described by thesauri terms and an entry from a hierarchical classification. The geographic scope is described by a standardized geographic term.

Within the database, 65% of the documents are books and papers, 25% are monographs, 5% yearbooks and 5% journals. English dominates the collection with 51% of the documents. Some 28% of the documents are written in German. French and Spanish are represented with 11% and 5%. Many other languages contribute a smaller amount of documents. All the descriptors assigned by the information specialists are in German. [5, 15]. For most non-German documents, the title has been translated into German. The heterogeneity in format, length and language represents a typical real world information problem.

The SWP developed 25 topics based on typical information needs on foreign policy. These topics represent the typical usage situation at the SWP. Political representatives of Germany need to be informed on a policy issue or a country. The SWP specialists query the database and create a set of relevant documents. More than 4000 documents needed to be assessed for relevance by these domain experts. A typical topic is: *Which factors determine the relation between China and the EU and the individual EU countries?* A set of some 20 topics can be considered to lead to reliable evaluation results [13].

4 Entry Vocabulary Module

An entry vocabulary module (EVM) aims at mapping between the user vocabulary and the terminology represented in the collection. The basic idea of the EVM is the mapping between full text terms and descriptors from the thesaurus [12, 16]. The relations between terms are calculated as probabilities derived from the coocurrence of the terms in documents [5, 12]. We will present the static version which is similar to classic term expansion and our dynamic version which is based on iterative integration of terms from different documents and fields. The stepwise expansion works without global knowledge and is consequently more efficient. The effectiveness can also be improved because a different document set is exploited for the term expansion. Classic term expansion cannot focus on the same documents and it cannot find the same terms. The iterative EVM is much more specific. The dynamic EVM also allows the integration of knowledge about the domain. The structure of the meta data can be used for optimizing the order of the EVM steps. We will show how the knowledge of a system engineer about the geographic terms can be integrated.

The base system applies Apache Lucene as a basic retrieval engine. All other components like the blind relevance feedback (BRF) module for the base run are based on a system which has been successfully evaluated at several CLEF campaigns [5]. The stopword list has been manually adapted to the data set. No language specific stemming was applied due to the high number of languages present in the data set. The query is also stripped of stop words and expanded with either the EVM or BRF.

4.1 Static Entry Vocabulary Module

According to an implementation by Gey et al., the EVM consists of four modules [13]:

- A data base sufficiently large to train the EVM
- A part-of-speech (POS) tagger to identify the nouns in the documents
- An algorithm to calculate the probabilistic relation between the two vocabularies based on the frequency of the coocurrence in documents.
- A basic retrieval system which adds the highly related terms, searches and presents results

An EVM is typically designed as a global system which considers the whole data set. However, recently local methods have been suggested. They consider only the documents in a set of documents retrieved in a first step. Comparative quality has been reported [16]. The influence of long documents can be a problem for an EVM and makes length normalization necessary. A local EVM does not necessarily require an normalization step because it an even out the influence of long documents by other means. Consequently, a local EVM is computationally much less demanding.

4.2 Dynamic Entry Vocabulary Module

We decided to adapt a dynamic approach which works in an iterative and cascading manner. One main advantage for the effectiveness of an EVM lies in the iterative functionality. After the initial retrieval step, the EVM can extract highly frequent thesauri terms of a certain categories and use them for a secondary retrieval step. Then, the documents in the result set can again be searched for highly frequent thesauri terms of another category until the final result is presented. The extracted terms are sent as queries to specific index fields. The dynamic EVM is also designed to bridge the language barrier. All intellectually added terms are in German, also for non German documents. That way, a German query can have hits in other languages and expansion can add terms in foreign languages which may lead to further hits.

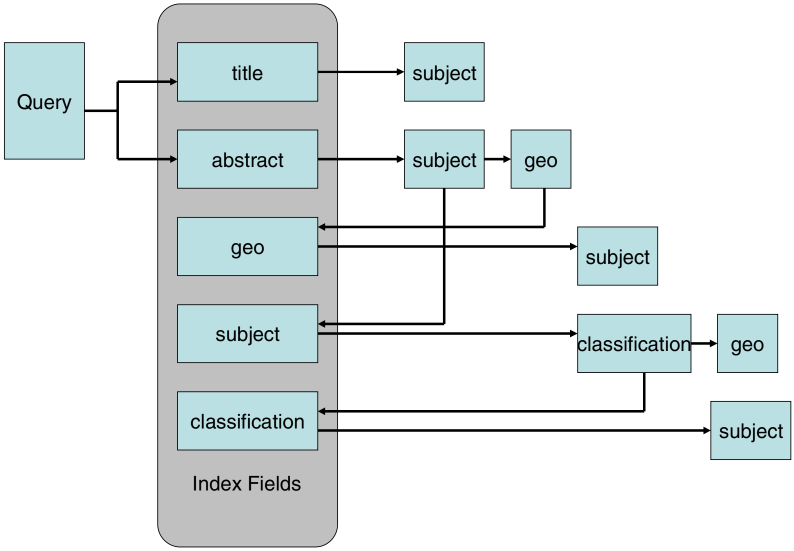


Fig. 1. Cascading Dynamic Entry Vocabulary Module

The order in which the categories are employed for the expansion is highly dependent on the domain. Step by step, the query is enriched based on both the meta data and the free text index. For the political science domain under consideration, the following cascading order was implemented:

- The query is sent to the title and abstract field and the subject (thesaurus terms) field of the returned documents is used for expansion.
- The query is sent to the abstract field only and the geographic entity field of the returned documents is used for expansion.
- The expanded geographic terms are sent as query to the geo field. The subject (thesaurus terms) field of the returned documents are used for expansion. This is supposed to select important terms for the geographic regions which are present in all queries.
- The subject terms are used as query and classification terms are extracted from returned documents.
- The subject terms are used as query and geographic terms are extracted from returned documents.
- A search with the classification terms retrieves subject terms.

The cascade is illustrated in figure 1. The parameters need to be carefully chosen in order not to finally select too many documents.

The effect of the EVM is illustrated by the following example. For the query East European Enlargement of the European Union, the top expanded terms are listed in table 1.

Table 1. Example for expansion terms

Term	Normalized Weight
European Union	1.000
Middle- and east European Countries	0.897
Enlargement of international actor	0.629
International regional political integration	0.539
Effect/ Cause	0.449
European Community	0.123
Community of Independent States	0.091
France	0.078
Russian Federation	0.072
Belorussia	0.063

5 Evaluation Results

In this chapter, the results of the runs are discussed. For all five runs, the first 200 documents were retrieved for each topic. These led to a set of 1000 documents for each topic. After elimination of multiple occurrences the whole set was given to assessors who performed the relevance judgments. All measures have been calculated with the trec_eval program version 7.0.

5.1 Evaluated Runs

Several parameters control the behavior of the EVM. We varied the number of documents considered and the number of terms extracted. Also the weights of the initial and the expanded query terms were modified. The following five runs were implemented:

- SwpBase1-Nmd: Search in title, abstract, no meta data, no EVM
- SwpBase1-Md: Search in title, abstract as well as in thesaurus terms and classification terms, no EVM
- SwpEvm1: Search in title, abstract as well as in thesaurus terms and classification terms, EVM adds six highest ranked terms from top 30 result documents.
- SwpEvm2: EVM adds six highest ranked terms from top 100 result documents. Weighting: expanded terms are weighted with their probabilistic relation score.
- SwpEvm3: EVM adds all terms over a cut-off from top 100 result documents. Weighting: initial query =10, expanded terms =1

5.2 Results

After pooling and relevance assessment, the performance of the runs was calculated. Standard information retrieval evaluation measures have been used [10]. The runs contributed to a pool which contained 2039 relevant documents of which 1018 were not in the query language German. The results are presented in table 2a and 2b and a recall-precision curve is shown in figure 2.

Table 2a. Performance of the five runs

Runs	Mean Average Precision (MAP)	Number of retrieved relevant docs.	Number of retrieved non-German relevant docs.
BaseRun1	0.0985	717	202
BaseRun2	0.1752	965	330
SwpEvm1	0.2138	1226	580
SwpEvm2	0.1980	1182	521
SwpEvm3	0.2046	1211	510

Table 2b. Performance of the five runs

Runs	Geometric Mean of Average Precision (GMAP)	Percentage retrieved relevant documents	Percentage retr. non-German relevant docs.
SwpBase1	0.0752	0.3516	0.1984
SwpBase2	0.1362	0.4733	0.3242
SwpEvm1	0.1881	0.6013	0.5697
SwpEvm2	0.1717	0.5797	0.5118
SwpEvm3	0.1884	0.5939	0.5010

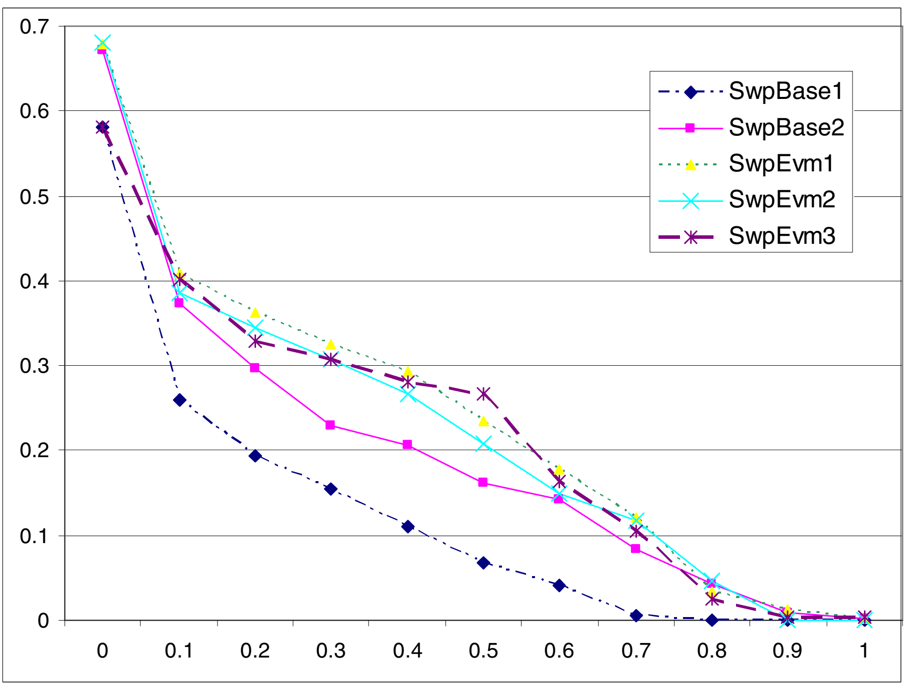


Fig. 2. Recall Precision Graph for all five Runs

6 Topic Analysis

Not all topics benefit from the EVM. As in most cases, the topic performance of the runs varies greatly. In order to explore the effects of the EVM on individual topics we adopted an innovative method suggested by Mizzaro & Robertson [9]. The matrix giving of topic and run performance was normalized. All average precision values in the matrix were normalized with the maximum and minimum performance of all systems for that topic. That means that a system which performs best for a topic receives the value 1 independent of the actual performance level of the topic. A visual analysis leads to three groups of topics. One group reaches a high performance level in the base runs and sometimes is even hurt by the EVM. These topics are shown in figure 3a. For these topics, the performance difference between the three EVM runs is high. Figure 3b shows a large number of topics which benefit from the EVM. These topics exhibit a low performance level for the two left runs and higher levels for the EVM runs. The performance differences between run EVM1 and EVM2 given in tables 1a and 1b are small. However, there is a large number of topics which are solved very well by EVM 1 and badly by EVM2 and vice versa. This cluster is shown in figure 3c. Note that some topics appear twice in figures 3a through 3c.

The performance differences were more closely analyzed to identify strengths and weaknesses of certain runs for types of topics. Typical topic features used in other analysis are vague and precise topics [14]. Defining types of topics leads to small sets of topics and consequently, such an analysis cannot lead to significant results.

Precise topics (What is the role of weapons of mass destruction and the fight against terrorism in transatlantic relationships?) seem to work well with base runs

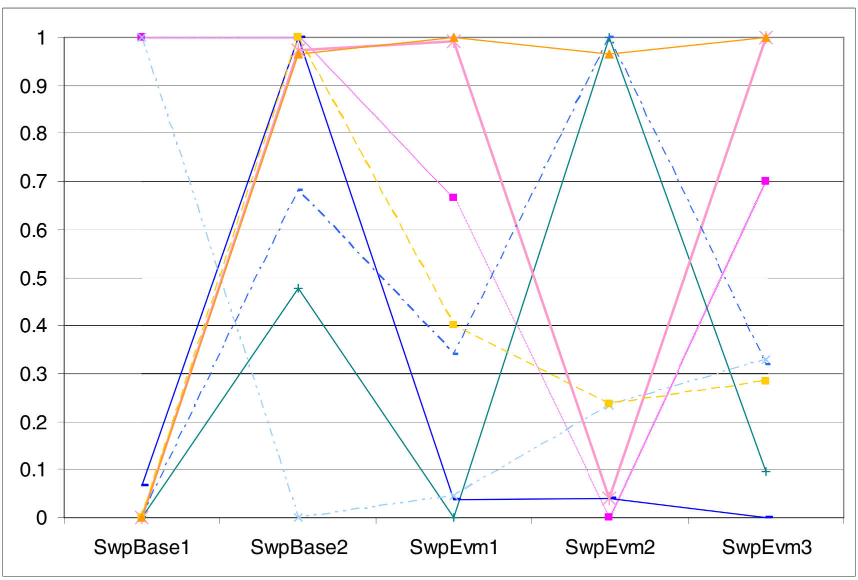


Fig. 3a. Topics with good base run results

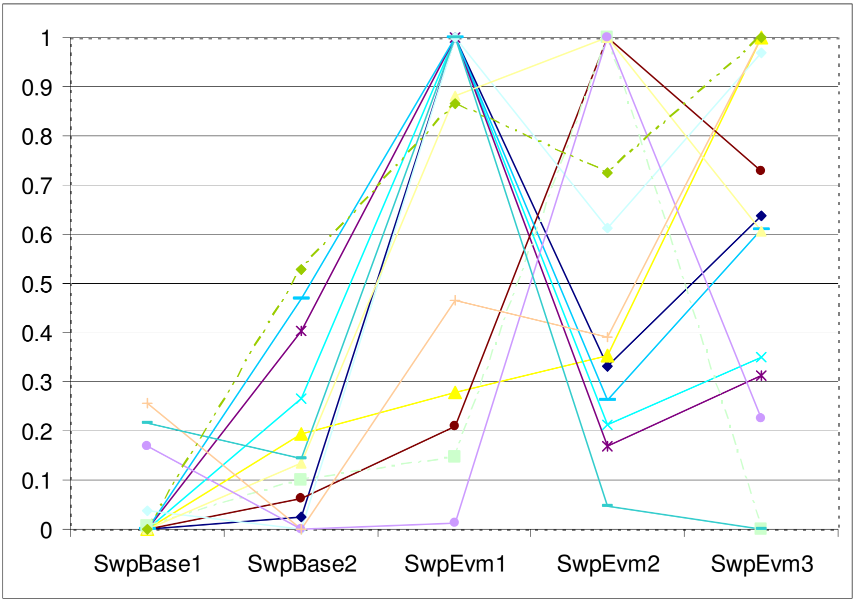


Fig. 3b. Topics which improved using the EVM

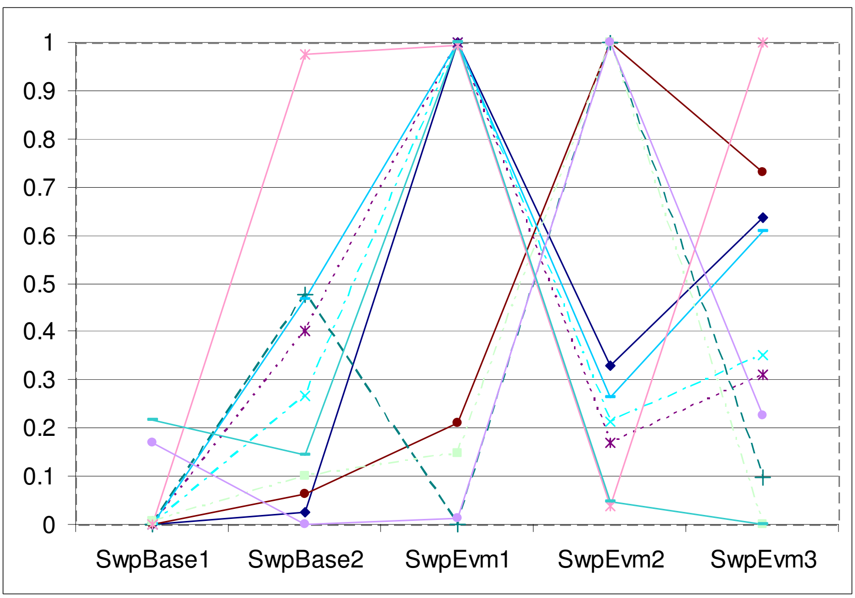


Fig. 3c. Topics with differences between EVM1 and EVM

whereas vague and open queries (How does the larger EU define its relation to European countries of the former Soviet Union and on the Western Balkan?) are solved better by EVM runs. This is especially true when a country name is mentioned in the topic (How did the relation between Nigeria and the USA develop in the past ten years?). Apart from that EVM runs work better for geographic entities different form a country (Which threats exist for maritime security in South East Asia?).

7 Conclusions

In our study, we presented a retrieval test for a collection of 600,000 political science documents. This test collection is domain specific, contains metadata and is multilingual. The collection is highly typical for real world business data.

For optimized retrieval, a dynamic entry vocabulary module (EVM) has been developed and evaluated with the retrieval test presented. A main contribution of the paper is a dynamic version of the EVM. The iterative algorithm of the dynamic EVM allows the integration of domain knowledge by assembling the steps in a different order. The iterative expansion of the query with additional terms leads to document optimally exploits the meta data a reaches documents which would not be accessed by standard term expansion techniques.

The results have been analyzed at the topic level with an innovative method which reveals weaknesses and strengths of the EVM for certain types of topics. Future information retrieval research will need to pay more attention to the topic specific optimization.

References

1. Braschler, M.: CLEF 2003 - Overview of Results. In: Evaluation of Cross-Language Information Retrieval Systems. Third Workshop of the Cross Language Evaluation Forum 2003. LNCS, Springer, Heidelberg (2004)
2. Buckland, M., Chen, A., Chen, H.-M., Kim, Y., Lam, B., Larson, R., Norgard, B., Purat, J., Gey, F.: Mapping Entry Vocabulary to Unfamiliar Metadata. In: Meta-Data 1999 Third IEEE Meta-Data Conference, Bethesda, USA (April 1999)
3. Buckley, C., Voorhees, E.: The Effect of Topic Set Size on Retrieval Experiment Error. In: Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR 2002), Tampere, Finland, August 11-15, 2002, pp. 316–323. ACM Press, New York (2002)
4. Fachinformationsverbund IBLK (German information Network) World Affairs Online (2006), http://www.fiv-iblk.de/eindex.htm and European Thesaurus on Internationale Relations and Area Studies online at, http://www.fachportal-iblk.de/index.php?&id=91&L=1
5. Hackl, R., Mandl, T.: Domain Specific Mono- and Bilingual English to German Retrieval Experiments with a Social Science Document Corpus. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 247–250. Springer, Heidelberg (2006)

6. Hellweg, H., Krause, J., Mandl, T., Marx, J., Müller, M., Mutschke, P., Strötgen, R.: Treatment of Semantic Heterogeneity in Information Retrieval. IZ-Arbeitsbericht Nr. 23, IZ Sozialwissenschaften, Bonn (2001), http://www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/index.htm#ab23
7. Hollink, V., Kamps, J., Monz, C., de Rijke, M.: Monolingual Document Retrieval for European Languages. Information Retrieval 7, 33–52 (2004)
8. Kluck, M., Stempfhuber, M.: Domain-Specific Track CLEF 2005: Overview of Results and Approaches, Remarks on the Assessment Analysis. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 212–221. Springer, Heidelberg (2006)
9. Mizzaro, S., Robertson, S.: HITS hits TREC – exploring IR evaluation results with network analysis. In: 30th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR), Amsterdam, pp. 479–486 (2007)
10. Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.): Evaluation of Cross-Language Information Retrieval Systems. Third Workshop of the Cross Language Evaluation Forum 2003, Trondheim, Norway, August 21-22, 2003 (2004)
11. Petras, V., Perleman, N., Gey, F.: Using Thesauri in Cross-Language Retrieval of German and French Indexed Collections. In: Advances in Cross-Language Information Retrieval, pp. 349–362 (2003)
12. Petras, V.: How OneWord Can Make all the Difference - Using Subject Metadata for Automatic Query Expansion and Reformulation. In: Working Notes for the CLEF 2005 Workshop, Vienna (September 2005)
13. Sanderson, M., Zobel, J.: Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. In: Proc. 28th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR), Salvador, Brazil, pp. 162–169 (2005)
14. Savoy, J.: Why do successful search systems fail for some topics? In: Proc. ACM Symposium on Applied Computing (SAC), Seoul, Korea, pp. 872–877 (2007)
15. Voorhees, E.: The TREC robust retrieval track. ACM SIGIR Forum 39(1), 11–20 (2005)
16. Xu, J., Croft, B.: Query Expansion Using Local and Global Document Analysis. In: Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 4–11 (1996)
17. Zobel, J.: How Reliable are the Results of Large-Scale Information Retrieval Experiments? In: Proceedings of the 21st Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR 1998), Melbourne, Australia, August 24-28, 1998, pp. 307–314. ACM Press, New York (1998)

Chinese Organization Entity Recognition and Association on Web Pages

Qi Zhang[1,*], Guopin Hu[2], and Lihua Yue[1]

[1] Department of Computer Science, University of Science, and Technology of China, Hefei, Anhui 230027
wizard@mail.ustc.edu.cn
[2] Department of EEIS, USTC, Hefei 230027, China

Abstract. In this paper, we consider the problem of automatic Chinese Named Entity Recognition (NER) on web pages and try to extract the association between recognized entities. Usually NER approaches mainly focus on plain text and get poor results on the Web pages of Internet. In this paper, we first explore the difference of plain texts and web pages for NER. Based on characteristic of HTML structure, we propose a set of unified methods to recognize and associate entities on web pages. In our experiments, the F-measure of organization name recognition is 73.6%, where 14.3% improvement is achieved beyond the baseline system. The F-measure of organization name-address association on page level is 77.5%, and the performance achieves 89.5% on corpus level, which indicates that our approach is quite effective and practical.

Keywords: Information extraction, NER, Data Integration, Chinese Named Entity Recognition.

1 Introduction

Named Entity Recognition (NER) is one of the key techniques in many applications such as Question Answering, Information Extraction, etc. On the other hand, with the rapid growth of the Internet, web pages have become a more and more important data source for information extraction. In this paper, we consider Chinese NER problem on the web pages.

Both Chinese and English NER have been widely studied [15] [16] [6] [12]. These researches mainly focused on plain text and did not take much consideration on HTML structures of web pages. The differences between NER on plain text and on web pages lie in: (1) there are separators (HTML tags) between phrases in web pages. Important entities are often separated by web page creators. (2) Web page has a hierarchical structure (DOM-Tree) which is very useful when trying to determine the association between recognized entities while it is hard to associate them in plain text especially when they are located far away from each other in context. (3) In Web pages, important entities, like organization name, will appear repeatedly in many different locations, such as title, Meta data or link anchor texts etc. It quite simplifies the recognition on Web pages.

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 12–23, 2008.

Based on the above analysis, we propose a set of unified methods for the Chinese NER problems on web pages. We only consider the problem of Chinese organization names and organization addresses recognition and association in this paper.

Our system can be presented in the following steps. First, given a web page, a DOM-Tree is built based on its HTML structures. The internal nodes of the tree are the HTML tags while the leaf nodes are the content texts. The text in the each leaf node is regarded as one entity candidate. We name this step as *Entity Candidate Generation*. Second, in *Candidate Recognition* step, we conduct Chinese organization name and address recognition algorithms on each entity candidate. Third, according to the structure of the DOM-Tree, our entity association algorithm is applied to determine the corresponding address for each organization name by some specified rules, which is named as *Entity Association* step.

We evaluate our algorithms on a large collection of web pages in which each individual NEs and their associations have been manually labeled. The experiment results show that our system can work efficiently on web pages.

The rest of the paper is organized as follows. We discuss the related work in next section, and in section 3, we present our entity candidate generation algorithms. In section 4 and 5 we proposed Chinese organization name and address recognition algorithm. We will discuss the association algorithm of these two entities in section 6. Section 7 shows our experimental setup and result. In last section, we conclude our work.

2 Related Work

English NER problems have been widely studied and impressive performance has been achieved. But for Chinese, NER is very different. There is no space character to indicate the word boundary and no standard definition of words for Chinese. The Chinese NE recognition and word segmentation are interactional in nature. Recently many methods have been proposed for Chinese NER.

Yu et al. [17] applied the HMM approach where the NER is formulated as a tagging problem. Their system requires information of POS tags, semantic tags and NE lists.

Jian Sun et al. [6] proposed an approach that integrates Chinese NER and word segmentation together using a class-based language model, and employs Viterbi search to select the global optimal solution. Unfortunately their model depends heavily on statistical information and can only be trained on large scale of labeled corpus.

Youzheng WU et al. [8] presented a hybrid algorithm that combines a class-based statistical model with various types of human knowledge. Human knowledge is incorporated to resolve data sparseness and other problems. Their experiments show that the performance of hybrid model increases remarkably against the original statistical model. For example, the precision and recall of Organization Names increase from 42.98% and 61.45% to 80.86% and 72.09% respectively.

A hybrid statistical model was proposed by Youzheng WU et al [12], which combines coarse particle features (Part-of-Speech model) with fine particle features (word model). The hybrid model overcomes the disadvantages of the word model and the POS model. They also incorporated human knowledge in this system to improve efficiency and effectively. F-measure of Organization Names on MET-2 is 80.21% in their experiment.

Wang and Shi [1] presented a simple rule based approach to organization name recognition in Chinese text. Their approach is based mainly on a list of heuristic rules. These rules help find left and right boundaries of a Chinese name entity and the possible composition of a name entity. In their test, the algorithm achieves an F-measure of 86.6% which is comparable to other statistical models but with less complexity and easier to implement.

But most of the aforementioned work deals with text only. To the author's best knowledge, there are no algorithms which aim the Chinese NER problem on web pages. In aforementioned work, HTML texts are always converted into plain texts in a pre-processing step. To process structured documents such as web pages these methods have to extract text out of web pages.

In this paper, we fully explore the structures of web pages and take advantage of these structures to help recognize Chinese named entities.

3 Entity Candidate Generation

The most difficult problem in Chinese NER is to determine the boundaries of the entities from whole text sentence. In plain text, there are hardly any cues for boundary identification. Boundary ambiguity is the main reason that leads to error recognition results. So previous NER algorithms always employed various methods to detect the boundaries, including statistic models and rule based ones.

However, the situation changes in web pages. As we know, web pages are created based on HTML structures. And HTML structure is constructed by tags, for example, *中国科技大学(Univ. of Sci. and Tech. of China)*. Therefore HTML tags are good cues that help us to detect the boundaries of Name Entities.

There are two observations we have made:

Observation 1: In web pages, important information units, such as organization name and address, are always emphasized with some particular tags (e.g. <title>, <b> and <meta> etc) or separated from other less important information with tags by the web page creators. Following is an example:

"*<TITLE>安徽中科大讯飞信息科技有限公司(ANHUI USTC iFLYTEK CO., LTD.)</TITLE><SPAN>地址：安徽省合肥市国家级高新技术产业开发区信息产业基地讯飞语音大厦(Address: iFlyTek Speech Building, State Level High-tech Industrial Development Zone, Hefei, Anhui) </SPAN>*."

Observation 2: Important entities, such as organization name and address always appear repeatedly in different locations, for example, in different positions of the same web page, in different pages of the same web site or even in different pages of different sites. Therefore the opportunity to extract correct named entity will be greatly promoted if the whole web site or internet is utilized. We can see the impact of this observation in our experiments.

Based on these two observations, we conclude that html tags can be regarded as the separators between entities. Observation 2 indicates that the *Recall* of our system will be acceptable even we fail to recognize all the entities in some particular web pages. It also indicates that recognition result validating is possible and critical in such applications. With result validating, some recognition error can be corrected by some other correct recognition from different positions of the same page or different pages.

Our entity candidate generation algorithm can be depicted in two steps as follows:

1. Parse the web page into a DOM-Tree in which the leaf nodes contain only text (without HTML tags).
2. For each leaf node, segment its text into phrases by punctuations. Each phrase will be regarded as a candidate entity and will be input into recognized algorithm later.

After entity candidate generation, a web page is converted into a DOM-Tree with phrases in its leaf nodes. Here we take a piece of HTML code from http://www.iflytek.com/contact as an example:

*“<P class=fonts><SPAN class=fonts1><IMG height=15 src="/images/mail2.gif" width=18>北京分公司:</SPAN>
联系人:李展眉
电话:010-62140301
传真:010-62140301
电子邮件:<A class = article href = " mailto: zmli@iflytek.com "><FONT color=#333333>zmli@iflytek.com </FONT> </A>
地址: 北京中关村南大街34号中关村科技发展大厦C座2509室
 邮编:100081 <STRONG> </STRONG></P>”*

Table 1. Generated entity candidates and their corresponding types

Candidates	Entity Type
北京分公司 (Beijing Branch)	Organization Name
联系人 (Contact)	—
李展眉 (Zhanmei Li)	Contact
电话 (TEL)	—
010-62140301	Phone Number
传真 (FAX)	—
010-62140301	Fax Number
电子邮件 (Email)	—
zmli@iflytek.com	Email Address
地址 (Address)	—
北京中关村南大街34号中村科技发展大厦C座2509室 (Room 2509, ZhongguancunScience and Technology Development Building C, 34 Zhongguancun South Street, Beijing)	Organization Address
邮编 (Zip Code)	—
100081	Zip Code

The entity candidates generated from the above piece of HTML code are listed in Table 1 with their corresponding types. From table 1, we can see that each candidate is either some type of entity we interest in or useless information.

Now we can conclude the key difference between our method and other Chinese NER algorithms: our method takes HTML structures into consideration when identifying the boundaries of entities. And in later recognition steps, we do not employ any extra processing (e.g., statistical models) to detect the boundaries. The candidate we get from *Entity Candidate Generation* step will be determined to be an entire named entity string or useless information later. In other words, we turn the problem of "where is the NE" into "whether a candidate is a NE," which simplify the recognition processing.

4 Organization Name Recognition

After candidate generation step, the leaf nodes in the DOM-Tree are segmented into candidate phrases. E.g., "中国科学技术大学(Univ. of Sci. and Tech. of China)", "北京中关村南大街34号中关村科技发展大厦C座2509室*(Room 2509, Zhongguancun Science and Technology Development Building C, 34 Zhongguancun South Street, Beijing)*", etc. The rest of our recognition work is to determine the type of each candidate entity.

In this paper, we are mainly interested in two types of named entities: Chinese Organization Name (CON) and Chinese Organization Address (COA). Some other entities, such as phone number, fax number, email address and zip code are quite easy to identify if proper lexicons and patterns are developed. Therefore we do not discuss them in this paper.

CON is a special kind of Chinese Name Entity similar to person name, which is impossible to enroll them into a limited list. Based on our observation, some rules are widely used in the naming of Chinese organizations:

1. Geographic prefix word is always chosen as the beginning of the name

For example, "北京饭店 (Beijing Hotel)" has the geo-prefix "北京 (Beijing)" which infers the hotel is in Beijing city. The geo-prefix words can be utilized in NER in two aspects:

a. Identify whether a candidate entity is a valid CON by checking whether there is a geo-prefix in it;
b. Identify the entity's geographic information which is very important in CON-COA association. Details will be presented in section 6.

2. For the ending of organization name, a suffix word is always chosen according the organization type

Take the example of "北京饭店 (Beijing Hotel)" again. The suffix "饭店 (Hotel)" informs us that it is a hotel. Almost all the CONs have suffixes except abbreviations. Suffix is useful in two reasons:

a. CON creators always try to choose the right suffix for its name in order to avoid any confusion for people. Therefore such suffixes almost exist in each full CON;
b. Organization types can be enumerated, therefore the choice of suffix is almost limited, e.g. "**公司** (Company)", "饭**店** (Restaurant)", "**大学** (University)".

Name suffix is the most important cue we have used in our recognition algorithm.

3. There are some words that almost impossible to be used in a CON

Some "Stop Words," such as "失败 (failure)" and "其它 (others)" will never be selected to be part of the name according Chinese naming habits.

There are mainly two categories of recognition algorithms: statistical model based and rules based. Statistic model need large scale of training data and the performance cannot be guaranteed. In our system, we employ the simple rule-based approach proposed by Wang and Shi [1] with a little modification. The difference between us is that there is no left and right boundary finding process in our methods. Because every phrase generated in entity candidate generation progress can be seen as a candidate named entity

5 Organization Address Recognition

Comparing to CON recognition, Chinese Organization Address (COA) recognition is quite more difficult because of the irregularities of COA naming rules and the randomness of abbreviations, especially in web pages.

There are three types of COAs that commonly exist in web pages:

A. Regular addresses which contain levels of administrative district

For example, "安徽省合肥市金寨路96号 (No. 96, Jinzhai Road, Hefei City, Anhui Province) ", or some abbreviation like "南京路311号 (No. 311, Nanjing Road) ". As we can see, there are hierarchical administrative district units in this type of addresses, such as "省(Province)", "市(City)", "路(Road)" and "号(No.)" in the first example.

B. Using building name as the address

For example, "黄山大厦1层 (1^{st} floor of Huang Shan Building) ". This type of addresses exists when the building is famous in local area and the detailed address can be ignored.

C. Using relative position to some organizations as the address

For example, "家乐福对面 (In the opposite of Carrefour)". In some abbreviation cases, people would like to use the position of some well-known buildings or companies to infer that of some other unknown ones.

We propose different strategies to recognize these three types of addresses.

For COA of type A, we employ Supported Vector Machine (SVM), which has been found quite effective for text categorization problems [13], to learn whether an entity candidate is a COA. After a comprehensive study, we formulate the features that are important for COA recognition as follows:

Binary features: There are totally 44 binary features in two classes: *Administrative District Unit* and *Special Word*. Some sample features of these two types are listed in Table 2. If an input address contains one of the *Administrative Region Unit*, e.g., "省 (Province)", the corresponding label of the feature will be set to TRUE, FALSE otherwise. If there is a *Special Word* in the input address and the word before it is a number, its label will be set to be TRUE, FALSE otherwise.

Inverse Document Frequency: In Chinese address naming habits, there are some words which are commonly used in addresses. For example, “解放”(liberation), “长江”(Yangtze River), there are several cities which have a road named “解放路”. So, we think it is useful to take additional features to present these common words. Here we use *Inverse Document Frequency* as the features. Document here means a candidate entity. We extracted 594 top frequency words from our training addresses. $IDF(w_i)$ can be calculated from the *document frequency* $DF(w_i)$, which is the number of documents that word w_i has occurred in. n is the total number of documents in the training data.

$$IDF(w_i) = \log\left(\frac{n}{DF(w_i)}\right) \tag{1}$$

Finally, we construct a feature vector of 638 dimensions for the classifier. The combination of binary features and IDF features ensure the performance of our SVM classifier.

Table 2. Some binary features for SVM

Administrative Region Unit	"省"(Province), "市"(City), "区"(District), "县"(County), "镇"(Town), "乡"(Town), "里"(Village), "村(Village), "弄"(Road), "巷"(Road), "街"(Street), "道"(Road), "路"(Road), "胡同"(Road), etc
Special Word	"号"(No.), "室"(Room), "层"(Layer) , "楼"(Floor) etc.

Table 3. Sample features used for COA of type B and C

Indicative Words	"地处", "位于", "地址", "座落" etc. (All means “Located in”)
Positional Adverbs	“附近”(Near), “前面”(In front of),“后面”(Behind), etc

The recognition of COA of type B and C are relatively more difficult. There are always building names or organization names in such addresses. So we can not apply *Machine Learning Methods* directly on these addresses. In the first step, we take the CON recognition methods proposed in section 4 to discover the building and organization names. And then we can only check if there are some *Indicative Words* or *Positional Adverbs* before or after the recognized CON. Some *Indicative Words* or *Positional Adverbs* words are listed in table 3. And the recognition process can be summarized as the following rules:

1. Apply the CON recognition procedure on the candidate entities.
2. If there is an *Indicative Word* in the beginning of the candidate and the rest of the candidate is an organization name, the candiate is determined as an COA of type B.
3. If there are both organization name and *Positional Adverb* in the candidate, it will be determined as a COA of type C.

Additionally, for all the COA types, some handcraft rules are applied, for example, the candidate should be shorter than some length threshold and should not contain any *Stop Words*.

6 Entity Association

In above sections, we propose algorithms to recognize Chinese organization names and addresses. But in practical applications, only complete organization information together with its name and address is meaningful or useful. The individual name or address is somewhat meaningless.

After recognition processing, all the name entity candidates in the leaf nodes of the DOM-Tree are determined. And our association processing step is conducted on the DOM-Tree.

Fig.1 shows a branch of a DOM-Tree example, where N_1 and N_2 are two CONs while A_1 and A_2 are two COAs.

There are two principle rules in our entity association processing:

1. **The nearer two entities located in DOM Tree, the higher possibility they will be associated with each other**

For example, in Fig1, the possibility of associating N1 with A1 is much higher than that of the associating between N1 and A2.

2. **No conflict exists in the geographical scope of two associated entities**

Take the example in Fig.1 again. The association between A_1 and N_2 is invalid according to this rule, because the geographical scope of A_1 is "浙江(Zhejiang Province)" while that of N_2 is "安徽(Anhui Province)".

Here we briefly describe our association algorithm:

Step 1: For each COA in leaf node, search backward (from right to left) until it finds the first CON that satisfied rule 2 or it meets the start the text. If the two entities are associated, store them in the local node and pass all the unassociated entities to their parent. Go to step 2.
Step 2: In internal nodes, receive all the unassociated entities from its children from left to right and insert them into a text string orderly. Associate these entities like that in step 1 and pass all the unassociated entities to their parents. Go to step 3.
Step 3: If it is not the root node, go to step 2. Stop otherwise.

We only present this simple association algorithm in this paper and we will study it in our future work.

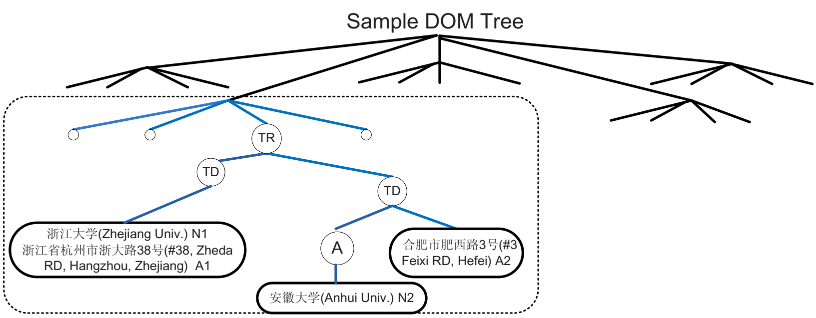


Fig. 1. Distance metric in DOM Tree

7 Experiments

7.1 Baseline System for Chinese Organization Name Recognition

In our baseline system, we implement the algorithm proposed by Wang and Shi [1] with slight modification. In the baseline system, we do not adopt POS-tags but we employ three lexicons to identify the left boundary, the right boundary and the constraints to conduct CON recognition.

These lexicons are common prefix list, common suffix list and common stop word list. In order to get these lexicons, we crawled 4000 name entities from http://yp.baidu.com/ with addresses. Each of these name entities will be segmented into words.

For each name entity, if its first word is a geo-word, a number, English word or a word which repeated more than 50 times in the training data will be put into the prefix list. And if its last word repeated more than 50 times will be considered as a suffix word.

7.2 Datasets and Lexicons

As for test data set, we collected 500 Chinese web pages, and each page contains at least one organization name or address. The URLs of these pages are collected through Google search engine with keywords such as "公司" (Company), and "饭店" (Hotel). In our test data, there are 3138 addresses and 5777 organization names. But there are actually 2891 Chinese Organization Entities in each of which there are at least one name and one address.

In the Chinese word segmentation process, we use a Chinese word dictionary with 64749 words. In the name identification, we employed 600 common organization name suffix, e.g., "公司 (Company)", "银行 (Bank)", and "饭店 (Hotel)", as the suffix dictionary.

In both organization entity recognition and association processes, a *Stop Word* list and a *Gazetteer* are needed. The *Stop Word* list has 1187 items and these items were impossible to be used in both names and addresses. And the *Gazetteer* contains a five-level hierarchical tree. A full path from root to the leaf can be formulated as "Country - Province – City – County - Town", e.g. "中国(China) – 浙江省(Zhejiang Province) – 杭州市(Hangzhou city) – 余杭区(Yuhang District) – 径山镇(Jingshan Town)."

7.3 Evaluation Measures for Recognition

We employ *Precision*, *Recall* and *F-measure* to evaluate our entity recognition results. In the evaluation, if the extracted entities (name or address) can exactly match to the annotated ones, we regard it as a correct extraction.

7.4 Experiments on Chinese Organization Name Recognition

In this experiment, we compared our system with the baseline system over the 500 HTML pages. For each web page of our test data, we first parse it into a DOM-Tree and conduct our recognition and association procedures on it. On the other hand, we parse the web page into plain text by removing all its tags for the baseline system.

Note that in the processing of our system, not only the entities highly suggested by the HTML tags but also the ones contained in the content will be considered as candidate ones. The results are shown in Table 4. We achieve of 14.3% improvement in F-measure than the baseline system. The reason that the performance of the baseline system is worse than that reported by Wang and Shi [1] is mainly because of the difference of the data set. The named entities in People Daily and MET-2 are more regular than those in web pages.

The precision of our system is not very high, because there are many candidate entities in the content which are not enclosed by HTML tags. But as we stated in observation 2, the same entity will distribute in many locations in the same page, for example, the title, Meta data, etc. We take one page from our test data as an example ("http://www.sunnychina.com/hotel/hotel_287.html"). In this page, the entity "合肥古井假日酒店 (Hefei Holiday Inn)" repeats 13 times in different locations, one in title, five in the Meta, etc. Eight of these thirteen existences are separated by HTML tags and punctuations which mean they can be recognized by our system. So the recall is quite acceptable.

7.5 Experiments on Chinese Organization Address Recognition

As we stated in the section 5, we adopt SVM for address classification. In order to get the best classification performance, we implement our approach using several SVM variations to select the best algorithm and parameters. We employ various kernel settings in SVM to test the address classification over out test data.

In the experiment, we choose 3230 labeled addresses (Downloaded from the yellow pages as stated in section 7.1) for training (1718 positive and 1512 negative) and 1697 for testing (825 positive and 872 negative).

Table 4. The results of our system and the baseline system

	Precision	Recall	F-measure
Baseline System	44.3%	74.2%	59.3%
Our System	60.2%	86.9%	73.6%

Table 5. The Comparison between Different Kernel Settings in SVM Training

	Precision	Recall	F-Measure
Linear	92.76%	94.67%	93.72%
Polynomial	48.73%	99.99%	74.36%
RBF	90.49%	93.45%	91.97%
Sigmoid	94.30%	78.18%	86.24%

Table 5 shows the experimental results. For the features we specified in section 5, linear kernel SVM achieves the best performance and is adopted for our system.

We conduct the SVM classifier over our 500 HTML pages. The precision of our classifier is 84.70%, the recall is 91.50% and the F-measure is 88.10%. The results show that our classifier can effectively recognize the Chinese organization addresses.

7.6 Experiments on Entity Association

In entity association experiments, we conduct two kinds of experiments: page level and corpus level. In page level experiments, two entities are regarded as correctly associated in one page when they are labeled associated in this page. And in corpus level experiments, the association of two entities is regarded as correct when they are labeled associated in any one of the pages in the corpus.

The reason to conduct corpus level experiments is that errors in the page level computing can be corrected by cross-validation in the page level results.

There are various ways to do cross-validating, and in this paper, the weight of each entity pair is computed according to the number of its occurrence. For example, if there are two different addresses A_1 and A_2 associated to the same name N_1, we count the number of the occurrence of pair (N_1, A_1) and pair (N_1, A_2). If the number of occurrences of pair (N_1, A_1) is larger than the other, we think the first one is more believable and make it as the correct answer.

The experiment results are listed in Table 6:

Table 6. The results of entity association experiments in page level and corpus level

	Precision	Recall	F-Measure
Page Level	69.10%	85.80%	77.50%
Corpus Level	88.50%	90.50%	89.50%

8 Conclusion

In this paper, we investigate the problem of Chinese organization names and addresses recognition from web pages. We first analyze the difference of the NER over plain text with that over web pages. Based on this analysis, we proposed an entity candidate generation method which converts the content of a web page into a list of candidate entities. The recognition processes are applied over these candidates to judge their corresponding types. After recognition, the entity association approach is applied on the individual entities. In our experiments, the F-measure of organization name recognition is 73.6%, where 14.3% improvement is achieved beyond the baseline system. The F-measure of organization name-address association on page level is 77.5%, and the performance achieves 89.5% on corpus level, which indicates that our approach is quite effective and practical.

References

1. Houfeng, W., Wuguang, S.: A Simple Rule-Based Approach to Organization Name Recognition in Chinese Text. In: Conference on Intelligent Text Processing and Computational Linguistics 2005, Mexico City, Mexico, pp. 769–772 (2005)
2. Leong, C.H., Ng, H.T.: Named Entity Recognition: A Maximum Entropy Approach Using Global Information. In: Proceedings of Coling 2002, Taipei, pp. 190–197 (2002)
3. Brown, P.F., DellaPietra, V.J., deSouza, P.V., Lai, J.C., Mercer, R.L.: Class-based n-gram models of natural language. Computational Linguistics 18(4), 468–479 (1992)
4. Chen, K.-J., Chen, C.: Knowledge Extraction for Identification of Chinese Organization Names. In: Proceedings of ACL work shop on Chinese Language Processing, pp. 15–23 (2000)

5. Zhang, Y., Huang, D., Zhang, L., Yang, Y.: Identification of Chinese Organization Names based on Statistics and Rules. In: Proceedings of JSCL 2001(Natural Language Understanding and Machine Translation), China, pp. 233–239 (2001)
6. Sun, J., Gao, J., Zhang, L., Zhou, M., Huang, C.: Chinese Named Entity Identification Using Class-based Language Model. In: Proc. Of the 19th International Conference on Computational Linguistics, Taipei, pp. 967–973 (2002)
7. Yu, H., Zhang, H., Liu, Q.: Recognition of Chinese Organization Name based Role Tagging. In: Proceedings of Advances in Computation of Oriental Languages, Beijing, pp. 79–87 (2003)
8. Wu, Y., Zhao, J., Xu, B.: Chinese Named Entity Recognition Combining a Statistical Model with Human Knowledge. In: The Workshop attached with 41st ACL for Multilingual and Mix-language Named Entity Recognition, Sappora, Japan, pp. 65–72 (2003)
9. Tsai, T.-H., et al.: Mencius: A Chinese Named Entity Recognizer Using the Maximum Entropy-based Hybrid Model. Computational Linguistics and Chinese Language Processing 9(1), 65–82 (2004)
10. Zhang, H.-P., et al.: Chinese Named Entity Recognition Using Role Model. Computational Linguistics and Chinese Language Processing 8(2), 29–60 (2003)
11. Callan, J., Mitamura, T.: Knowledge-Based Extraction of Named Entities. In: CIKM 2002, McLean, Virginia, USA (2002)
12. Wu, Y., et al.: Chinese Named Entity Recognition Based on Multiple Features. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, pp. 427–434 (2005)
13. Hearst, M.A.: Trends and controversies: support vector machines. IEEE Intelligent Systems 13(4), 18–28 (1998)
14. Hill, L.L., Frew, J., Zheng, Q.: Place names: the implementation of a gazetteer in a georeferenced digital library. Digital Library 5(1) (January 1999)
15. Chinchor, N.A.: Overview of MUC-7/MET-2. In: Proceedings of the Seventh Message Understanding Conference (MUC-7) (April 1998)
16. Xun, E., Huang, C., Zhou, M.: A Unified Statistical Model for the Identification of English BaseNP. In: Proceedings of ACL 2000, Hong Kong (2000)
17. Yu, S.H., Bai, S.H., Wu, P.: Description of the Kent Ridge Digital Labs System Used for MUC-7. In: Proceedings of the Seventh Message Understanding Conference (1998)
18. Collins, M.: Ranking Algorithms for Named Entity Extraction: Boosting and the Voted Perceptron. In: Proceeding of ACL 2002, pp. 489–496 (2002)
19. GuoDong, Z., Jain, S.: Named Entity Recognition Using a HMM-based Chunk Tagger. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), Philadelphia, July 2002, pp. 473–480 (2002)

Language Model Based Temporal Information Indexing

Andrzej Bassara

Department of Information Systems,
Poznan University of Economics,
ul. Niepodleglosci 10, Poznan, Poland
A.Bassara@kie.ae.poznan.pl

Abstract. Temporal information indexing is a specialization of general information indexing, which aims at creation of: clear, concise, correct and comprehensive representations of document important features, with respect to possible information needs. In case of temporal indexing, temporal features represent so called, time horizon or valid time of information, i.e. time period, to which information presented in a document relates.

Approaches to temporal information indexing are based mainly on temporal expressions extracted from document contents. This method, although satisfactory for many applications, is very limited. Temporal expressions are only one of many means of relating information to time used in written language.

In this paper, the possibility of temporal information indexing based on other, more broad and frequent document temporal feature is evaluated. Soundness and effectiveness of temporal indexing based on document syntactic features is presented.

Keywords: information retrieval, temporal information retrieval, temporal expressions, machine learning, indexing, temporal indexing.

1 Introduction

Most of Information Retrieval (IR) Systems provide a functionality of searching for documents, which fulfill certain syntactic criteria. Users may specify their information needs in form of queries, which are usually formulated as sets of words or phrases, called query terms, that should to or should not appear in documents that are to be treated as relevant ones. Often, query terms may be connected by Boolean operators [4].

This approach is based on algorithmic definition of relevance. Clearly, the relevance of document to the information is not only dependent on documents and queries themselves but also on additional factors [15,9], like: user knowledge, tasks he is performing or device he is using, just to name the few. However for the sake of simplicity, we are not concerned with them. This work is focused only on the problem of indexing, which can be examined in reasonable separation from the rest of IR components.

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 24–35, 2008.

Bearing in mind this simplification, a comparison of query terms against documents (in fact documents representations known as indexes), is executed at the syntactic level and may be performed, for instance, through string matching. This approach is, however, limited by problems associated with terms ambiguity. Each concept may have more than one syntactic representation (is denoted by more than one term) and each term may denote many concepts. To solve this problem, comparison of query and index terms may be performed on stemmed or lemmatized words. The process is sometimes supported by thesauruses or synonyms dictionaries, or even performed on ontological level.

In terms of Herbert Simon terminology, this approach is „satisficing“. Although, neither optimal nor powerful enough to solve many types of queries, it is satisfying and sufficient for most common types of information needs and is therefore successfully applied in many commercial solutions.

Table 1. Sample query with temporal criteria

Document:	The board of the Globe Trade informs that during 16 August 2006 ...
Information need:	all documents that relate to the third quarter of the last year
Query:	the third quarter of the last year

Unfortunately, the above approaches are not sufficient for solving queries containing temporal criteria. Table 1 presents a sample information need, a query and a document. It appears that the query and the document are not syntactically related (do not share any terms). Semantic comparison based on synonyms or concepts comparison will also yield null relevance. However, the document seems to be, at least partially, relevant while *16 August 2006* is part of *the third quarter of the last year*.

This limitation is caused mainly by an indexing process. In Information Retrieval, a document index is a simplified and machine processable representation of document important features. This representation substitutes the actual document content in the retrieval process. Relevance of a document to a given query is than based not on comparison of the query against the document but the query against the document index.

Features, that have been traditionally assumed to be important, are only words (sometimes stemmed or lemmatized) derived from document content. They are often accompanied by their position in the indexed document as well as weights representing their importance.

1.1 Problem Statement

To overcome the highlighted limitation a new specialized index must to be created that will represent properly document temporal features. The index should be supported by formalized model of time, which defines: basic time units (intervals, time points, granules etc.) [6,22], relations that may hold between them [3,8] and basic time/calendar arithmetic [17]. In this case, an indexing term may be simply defined as a time model element, for instance as a time point or a granule.

This approach allows to perform a comparison of a query against documents at the level of selected time model. As a consequence, not syntactic, but temporal relations between terms may be utilized (like inclusion, *16 August 2006* happened during *the third quarter of the last year*).

The major challenge in this approach is proper construction of the document index. The value of temporal index may be set by human expert [11]. This method is however out of question for large document sets. In most cases indexing process needs to be conducted fully automatically.

Often it is proposed to use temporal references extracted from document content [1,12,10]. It is assumed, that if temporal expression appears in the document, then the document is somehow related to the date/span denoted by this expression. There exist a variety of types of temporal expression that may be utilized for that purpose (see [7,13,19] for details).

Unfortunately, this approach, besides the fact that is computationally expensive, has one major drawback. It is constrained to the very limited set of temporal features used to temporarily anchor information. Temporal expressions are direct way of expressing temporal information („something happened at particular date“) and due to their relative unambiguity are often utilized in applications that requires this characteristic – like in news articles, medical records or financial statements.

However, as Mani and Wilson [14] point out, in many cases, the more usable way of information temporal anchoring is relating information to known states or events („something happened during/before/after/on some event“). Each state/event is characterized by specific dates (dates of occurrence, start, end etc.). Dates specific for events presented in a document may be than included in its temporal index. For instance, expression „during Great Depression" may be included in temporal index as a span 1929-1933.

This approach is complex for two reasons. It requires the ability to extract events, which is reasonably easy for named events (like beforementioned „Great Depression“) but may be very complex task for unnamed events which are specific for certain community („when Company X entered the market"). Moreover, this approach requires knowledge on dates specific for events that appear in the document.

Therefore, we have decided to use even more indirect temporal features – language model. We assume that the probabilistic language model is time dependent, i.e. time model will vary among documents related to different time periods.

1.2 Related Work

The problem of searching for information using temporal criteria is vital, especially in domains where time aspect plays a crucial role (for sample applications see [2,11]). In financial markets analysis, which is in the center of our interest, users combine: transaction data, financial statements and other textual information (including news stories) to form complex view on a market situation. To perform automatic combination of information from these sources at lest two dimensions must be considered: subject (market, market segment, company etc.) and time period. Subject and time are integral characteristics of financial statements and transaction data but not news stories. This requires documents to be indexed both by subject and by time to which information contained in document relate.

Unfortunately, the issue of temporal indexing, haven't brought much attention in the literature. There are not many publications in this area and presented solutions relay exclusively on different types of temporal expressions [11,10,16].

There is, however, much research in the area of temporal information extraction and some in the area of text corpora analysis with temporal criteria. We were mainly inspired by the former, especially by works from [21,5].

Swan and Allan [21] analyzed change of phrases significance over time. Authors used statistical tools to extract phrases which are highly time dependent. Time dependent phrase, is a phrase which occurs significantly more frequently in documents published in specific period, then outside it.

Dalli et al. [5] proposed a method for automatic dating of documents based on their vocabulary. Authors claim, that some words exhibit seasonal characteristics, ie. the function of conditional probability that given word appears in a document with certain publication date posses seasonal characteristics. This knowledge may then be used in turn to assess publication date of document using exclusively its vocabulary.

2 Methods

To prove the soundness of our approach, following plan was executed:

- Reference Index Preparation – The solution needs to be properly evaluated. This can be achieved using objective evaluation metrics. These metrics require a reference index, i.e. an index that will be regarded as a baseline, against which our approach will be compared.
- Verification of Basic Assumptions – Our work is based on two assumptions. Firstly, we assume that semantically similar documents tend to have similar temporal indexes. Moreover, semantically similar documents should be syntactically similar. Therefore, syntactically similar documents have similar temporal indexes. Secondly, we assume that some terms are time dependent, what means that they appear more frequently in document related to certain periods, then in other documents. Prior to construction of the indexing mechanism, these assumptions had to be verified.
- Indexer Construction and Evaluation – Based on verified assumption the indexing method may be created and evaluated.

2.1 Reference Index Preparation

In order to properly evaluate the presented solution, a temporarily indexed corpus was needed. Unfortunately, such a corpus was unavailable, at least for Polish language, which was in the center of our interest. The corpus had to be constructed from scratch. More than 25000 of documents (385934 sentences with average 15,4 sentences per document) have been collected. The corpus consists of press articles from major Polish financial newspapers, which was published between years 2000 and 2007.

The corpus has been temporarily indexed. Due to its size, manual indexing was out of question, so it was decided to use an existing automated indexer. The indexer that uses temporal references extracted from text was employed. The created index is regarded as a reference index, and the indexing method as a reference method for the rest of presented work.

This approach has one major drawback. Method, which is inherently based on limited set of temporal features, was used as a reference method. As noted in the introductory section, temporal references are only one out of many means of expressing temporal information. Therefore, special care must be taken, when interpreting the results of performed experiments. They do not present absolute measures of effectiveness, but rather correlation with existing approaches.

Moreover, the reference indexer did not exhibit 100% effectiveness. The indexer utilizes manually crafted rules for extraction and normalization of temporal expressions. Just to shed some light on the level of error, the indexer was evaluated against manually annotated documents (more then 1000 of temporal expressions). The precision and recall of recognition was respectively: 0.9845 and 0.972. More then 91% of correctly extracted references, were normalized properly. Therefore, in total, 90.67% of temporal expressions that appeared in text were extracted and normalized properly.

Each document was than automatically indexed. Index was defined simply as a set of indexing terms. An indexing term is a pair (I, G), where I is a granule index within granularity level G (for terminology and calendar arithmetic see [17]). Granularity levels represent different level of abstraction used for formulating temporal expressions and are related to a specific calendar. Granularity levels may include: a day of the month (DAY, e.g. *,,12/07/2007‘‘*), a day of the week (DOW, e.g. *,,on Monday”*), a month of the year (MTH, e.g. *,,in January“*), a quarter of the year (QTR, e.g. *,,last quarter”*), a year (YER, e.g. *,,in 2007“*) etc. A granule index is a number which represents granule position in granules sequence. We treat a granule index as a number of granules between an analyzed granule and a reference granule. A reference granule is a granule indexed as 1. For days it is day 01-01-0001. For other granularities, these are granules which contain this day.

For instance, a document i from the corpus D which contains only two temporal references ,,year 2000” and ,,first quarter of 2000“ will be indexed as:

$$Idx_{ref}(\boldsymbol{d_i}) = \{(2000, YER), (8001, QTR)\} \tag{1}$$

2.2 Preliminary Results

Before the development of the new indexing method could have been started, basic assumption must have been verified. It was assumed that syntactically similar documents should have similar temporal indexes.

In order to compare documents syntactically Vector Space Model (VSM) was employed. Each document was represented as a vector of terms $\boldsymbol{d_i} = (w_1, \ldots, w_k)$, where w_l is a weight of term l in document i (weights were computed according to TF*IDF formula). A term was defined as a word or a noun phrase extracted from documents content. To solve the problem of rich morphology of Polish language, lemmatization technique was used (see [23]). Also too frequent or too rare terms were rejected from the model.

For each document in the corpora, syntactic similarity based temporal index (Idx_{syn}) was computed. The index was defined as a weighted average of reference temporal indexes (Idx_{ref}) of most similar documents. Therefore, a notion of term weight had to be introduced. Weight $wg_{type}(\boldsymbol{d_i}, I, G)$ is a function that for each document $\boldsymbol{d_i}$ and for each granule (I, G) in index (Idx_{type}) assigns value from 0 to $1 - wg_{type} : D \times I \times G \rightarrow [0, 1]$. In case of reference index, the value is 1 if granule is present in the index set otherwise 0.

The weight of granule I in granularity G for document $\boldsymbol{d_i}$ in syntactic similarity based index is than defined as:

$$wg_{syn}(\boldsymbol{d_i}, I, G) = \frac{\sum_{\boldsymbol{d_j} \in D}^{sim(\boldsymbol{d_j}, \boldsymbol{d_i}) \geq l} wg_{ref}(\boldsymbol{d_j}, I, G) * sim(\boldsymbol{d_j}, \boldsymbol{d_i})}{\sum_{\boldsymbol{d_j} \in D}^{sim(\boldsymbol{d_j}, \boldsymbol{d_i}) \geq l} sim(\boldsymbol{d_j}, \boldsymbol{d_i})} \tag{2}$$

where, $sim(\boldsymbol{d_j}, \boldsymbol{d_i}) \in [0, 1]$ is similarity between documents $\boldsymbol{d_j}$, $\boldsymbol{d_i}$, l is similarity threshold. For this experiment similarity was defined as a cosine of the angle between two document vectors.

The index is than defined as set of pairs (G, I) for which the value of weighted average is more or less k of maximum value (the value of k was assigned experimentally, please see language model based index for details on influence of k on precision and recall):

$$Idx_{syn}(\boldsymbol{d_i}) = \{(I, G) : wg_{syn}(\boldsymbol{d_i}, I, G) \geq k * max[wg_{syn}(\boldsymbol{d_i}, I, G)]\} \tag{3}$$

The effectiveness of such an indexer may be measured using precision and recall measures. Both reference index as well as syntactic similarity based index may contain many granules, therefore these measures are defined as:

$$precision = \frac{\sum_{\boldsymbol{d_j} \in D} |Idx_{syn}(\boldsymbol{d_i}) \cap Idx_{ref}(\boldsymbol{d_i})|}{\sum_{\boldsymbol{d_j} \in D} |Idx_{syn}(\boldsymbol{d_i})|} \tag{4}$$

$$recall = \frac{\sum_{\boldsymbol{d_j} \in D} |Idx_{syn}(\boldsymbol{d_i}) \cap Idx_{ref}(\boldsymbol{d_i})|}{\sum_{\boldsymbol{d_j} \in D} |Idx_{ref}(\boldsymbol{d_i})|} \tag{5}$$

Figure 1 depicts results of this experiment conducted for one granularity level only – for years. The results follows our expectations – the higher the similarity threshold (l) the higher is precision. Unfortunately, the probability that there exist in a collection at least one document whose similarity exceeds the given threshold drops drastically. Recall follows this pattern only to certain degree. If the similarity threshold is relatively low (lower then 0.5), the lower its value

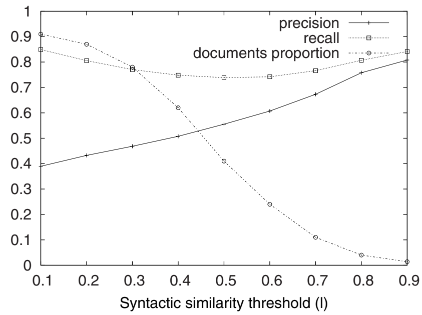


Fig. 1. Temporal indexing based on syntactic similarity

gets, the higher is recall. This happens because the index is based also on not very related documents and as a result, temporal index contains irrelevant terms, which may span almost the whole time axis.

It appears that if information on documents similarity is present in the collection and if at least part of documents is temporarily indexed, this information may be used directly in the temporal indexing process.

The second assumption was related to time dependence of terms. We believed that, at least for some terms, the probability that a given term appears in a document is dependent on the time to which the document relates, i.e. is dependent on its temporal index.

For each term and for each granule, conditional probability of a term occurrence in a document related to this granule was computed:

$$P(w_l|(I,G)) = \frac{|\{\boldsymbol{d_i} \in D : (I,G) \in Idx_{ref}(\boldsymbol{d_i}) \wedge \boldsymbol{d_i}(w_l) > 0\}|}{|\{\boldsymbol{d_i} \in D : (I,G) \in Idx_{ref}(\boldsymbol{d_i})\}|} \qquad (6)$$

where, $P(w_l|(I,G))$ is shorthand for $P(\boldsymbol{d_i}(w_l) > 0|(I,G) \in Idx_{ref}(\boldsymbol{d_i}))$.

It appears that there exist three classes of terms (Fig. 2):

seasonal terms (for instance ,,December") – These are terms, for which frequency of occurrence is periodically significantly higher. This group consists of terms related to cyclic events. In this case, it is reasonable to utilize seasonal analysis. It is however out of the scope of our work (see [5] for details).

time-dependent terms (for instance, ,,multiscreen.tv‘) – These are terms, which appears relatively frequently for limited number of granules, for other granules their frequency is significantly lower. This group consists of terms related to temporarily significant events.

time-independent terms (for instance ,,where‘) – These are the terms, which frequency does not exhibit any significant change over time.

Terms that constitute first two groups are significant for proposed method. Time-independent terms may be rejected and form temporal stoplist, i.e. a list of terms, which are irrelevant for assessing time index of the document. We have,

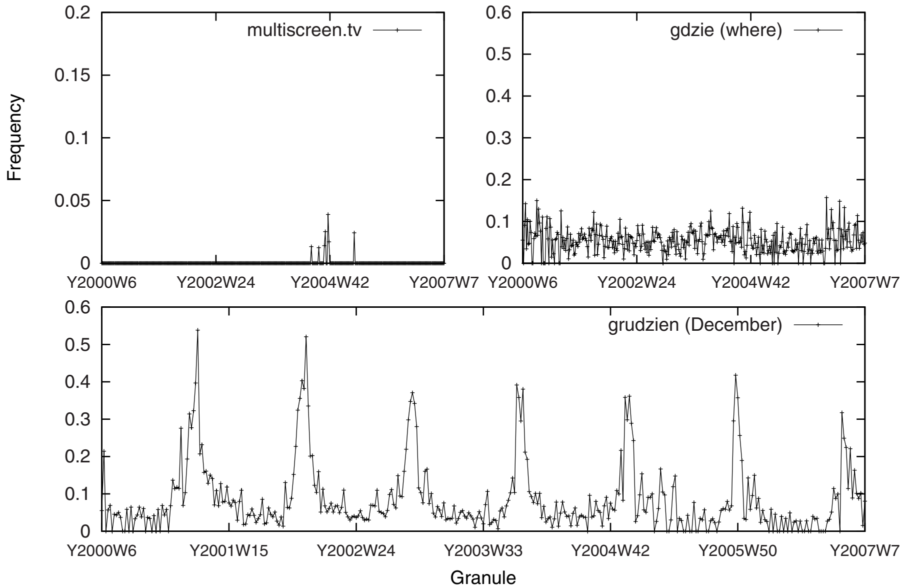


Fig. 2. Temporal indexing based on syntactic similarity

however, decided not to reject these terms, while their presence should not have significant, negative impact on the effectiveness of the developed indexer.

2.3 Classifier Construction

Temporal indexing, and indexing with controlled vocabulary in general, may be regarded as a specialization of classification problem. In this approach, each individual term is treated as a separate class. Classifier (actually a set of classifier) may then assign each document to a set of terms.

Formally, a text classification [20] is a task of assigning a binary value (fuzzy classification [18] may also be applied) to each pair $(\boldsymbol{d_i}, c_j) \in D \times C$, where C is a set of classes. The value T is set for a pair $(\boldsymbol{d_i}, c_j)$ in case, if decision is made to classify $\boldsymbol{d_i}$ to class c_j, otherwise F.

This formalization, however, does not reflect different effectiveness levels that may be achieved by various classifiers. A classifier may, and usually does, make some errors during classification process. Therefore, it is necessary to distinguish correct classification (reference classification) from results achieved by an analyzed classifier.

Classification task is therefore [20] defined as an approximation of target (reference) classification function $\Phi : D \times C \rightarrow \{T, F\}$ which defines desired, proper classification, with function $\phi : D \times C \rightarrow \{T, F\}$ called classifier. Both function should be as close as possible. The measure of similarity is called classifier effectiveness. In this work measures of: precision, recall and error level are used.

Proper construction of a classifier requires definition of three its aspects: features, classes and the actual model which stands behind the classifier.

The classifier does not directly process documents, but their representations. Guidelines for generic text classification/clustering were followed. Document representations for text classification, were exactly the same, as for syntactic similarity computations, what greatly simplified the process.

A classifier transforms documents features in a set of classes. In this case, a document class is a single granule (I, G). It can be easily noted, that there needs to be at least one classifier for each granularity level. For testing purposes, classifiers for years and for quarters were trained. The decision is dictated by the size of the available training corpora. While each of the classes, had to be represented by significant number of documents.

For granularity levels of years and quarters, naïve Bayesian classifiers were constructed. The choice of classification model was dictated by its effectiveness and form of input data. Actually, the learning process didn't have to be performed, while the conditional probability of terms occurrence for different granules were known, cause it was computed for preliminary experiments.

As it is assumed that document may be assigned to each granule with equal probability, the probability of document d_i occurring in granule (I, G) may be computed according to following formula (please note, that $P(w_l|(I,G))$ were computed using Laplace smoothing):

$$P((I,G)|\boldsymbol{d_i}) \propto \hat{P}((I,G)|\boldsymbol{d_i}) = \prod_{l=1}^{n} P(w_l|(I,G)) \tag{7}$$

An analyzed document is then classified to granules for which the value of probability is maximal or is more then $k * max$, where $k \in (0, 1]$.

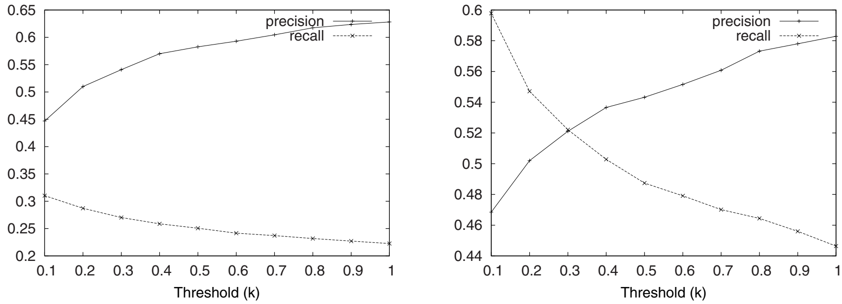
$$Idx_{lng}(\boldsymbol{d_i}) = \{(I,G) : \hat{P}((I,G)|\boldsymbol{d_i}) \geq k * max(\hat{P}((I,G)|\boldsymbol{d_i}))\} \tag{8}$$

3 Results

Each document in the testing set (10% of the corpus, which wasn't used for learning) was indexed by the developed method. In this process two classifiers were employed, one for granularity of years and one for quarters. Figure 3 presents effectiveness of both of them for different levels of similarity threshold k (see equation 8).

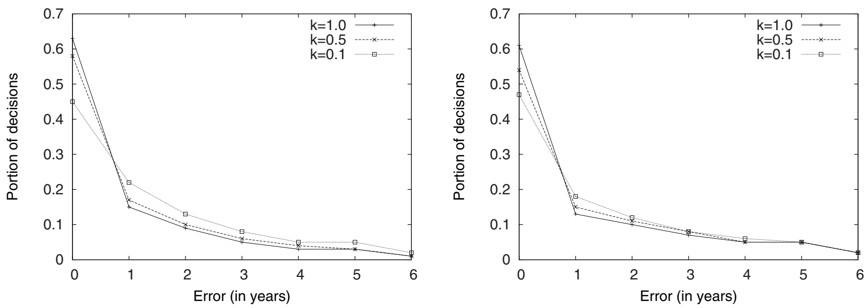
It appears that both of them achieve the precision of 60%, so around 40% of indexing terms, that were assigned to documents, were not present in the reference index. The actual problem is, however with the recall. The classifiers finds not all but only the most significant terms for each document (each document must be assigned to at least one granule). The problem will probably be more significant for finer granularity levels like months or days.

Contrary to most classifiers, temporal classifier may be evaluated by additional effectiveness measure. In most cases, correctness of classifier output may



(a) Quarter (b) Year

Fig. 3. Effectiveness of the classifier



(a) Quarter (b) Year

Fig. 4. Error of the classifier

be evaluated to binary value $(correct, incorrect)$. In case of temporal classifier, different degrees of errors may be committed. For example, if correct class for a document is $(2000, YER)$ and the classifier assigns $(2001, YER)$ the error is less severe as if $(1990, YER)$ were assigned.

Therefore, an error level were introduced. Error level defines the severity of an error in time units and is defined as a number of granules between class assigned by the classifier and the closest (in terms of calendar arithmetic) granule from the reference index for an analyzed document. If error level has value of 0 then the classification decision was correct. Figure 4 presents error levels for developed classifiers. Around 60% of classification decisions are correct, additional 20% have an error level of less or equal one year. The number of erroneous decisions drops with increasing error level.

4 Discussion

Presented results are surprising positive, while training and testing corpora spans only 8 years. Even for such a short period changes within language model are significantly enough, to consider their usage in the indexing process.

A particular care must be taken when interpretting these results. A reference index, which was used for learning and then for evaluation was based on limited set of temporal features. Therefore two factors may influence the results: weakness of the classifier and variance of two indexing methods.

Part of erroneous decisions may be a result of employed classification model. The model is based on all features, no particular attention was paid to features selection and terms which are time-independent was also used in the classification process. Some errors may arise from the reference index, which due to its limitation does not cover all temporal features of indexed document.

The method may be further developed. It may be improved by means of proper feature selection based on terms temporal characteristics. The shift of processing level from syntactic to semantic may also be considered.

If prestented results are to be generalized, the method requires broader evaluation. It needs to be evaluated against human annotated corpus or at least against corpus annotated by different indexing mechanism. It must be also checked, if presented findings hold for different categories of documents (natural language, subject or style).

This method may be used in practice as a supplementary technique for existing method of temporal indexing. It may be used to index documents, which do not contain any temporal references, the method may be also used to extend or verify temporal index created by other indexing method.

References

1. Abramowicz, W., Kalczynski, P., Kaczmarek, T., Wecel, K.: Time-indexer: a tool for extracting temporal references from text documents. In: The Information Resources Management Assciation International Conference 2003, pp. 832–835 (2003)
2. Abramowicz, W., Kalczynski, P.J., Wecel, K.: Filtering the Web to Feed Data Warehouses. Springer, New York, Inc., Secaucus, NJ, USA (2002)
3. Allen, J.F.: Maintaining knowledge about temporal intervals. Commun. ACM 26(11), 832–843 (1983)
4. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley, Reading (1999)
5. Dalli, A., Wilks, Y.: Automatic dating of documents and temporal text classification. In: Proceedings of the Workshop on Annotating and Reasoning about Time and Events. Association for Computational Linguistics, Sydney, Australia, July 2006, pp. 17–22 (2006)
6. Euzenat, J., Montanari, A.: Handbook of Temporal Reasoning in Artificial Intelligence. In: Time Granularity, Elsevier Science Inc., Amsterdam (2005)
7. Ferro, L., Gerber, L., Mani, I., Sundheim, B., Wilson, G.: TIDES 2005 Standard for the Annotation of Temporal Expressions (2005)
8. Freksa, C.: Temporal reasoning based on semi-intervals. Artif. Intell. 54(1-2), 199–227 (1992)
9. Ingwersen, P.: Information Retrieval Interaction. Taylor Graham (1992)
10. Kalczynski, P.J., Chou, A.: Temporal document retrieval model for business news archives. Inf. Process. Manage. 41(3), 635–650 (2005)

11. Llavori, R.B., Cabo, M.J.A., Barber, F.: Discovering temporal relationships in databases of newspapers. In: IEA/AIE 1998: Proceedings of the 11th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, London, UK, pp. 36–45. Springer, Heidelberg (1998)
12. Llido, D., Llavori, R.B., Cabo, M.J.A.: Extracting temporal references to assign document event-time periods. In: Mayr, H.C., Lazanský, J., Quirchmayr, G., Vogel, P. (eds.) DEXA 2001. LNCS, vol. 2113, pp. 62–71. Springer, Heidelberg (2001)
13. Mandel, M., Walker, C.: Time Annotation Guidelines For Less Commonly Taught Languages (2006)
14. Mani, I., Wilson, G.: Robust temporal processing of news. In: ACL 2000: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, USA, pp. 69–76 (2000)
15. Mizzaro, S.: How many relevances in information retrieval? Interacting with Computers 10(3), 303–320 (1998)
16. Nagypal, G.: Improving information retrieval effectiveness by using domain knowledge stored in ontologies. In: OTM Workshops, pp. 780–789 (2005)
17. Ning, P., Wang, X.S., Jajodia, S.: An algebraic representation of calendars. Annals of Mathematics and Artificial Intelligence 36(1-2), 5–38 (2002)
18. Roubos, H., Setnes, M., Abonyi, J.: Learning fuzzy classification rules from data. In: RASC conference (2000)
19. Sauri, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., Pustejovsky, J.: TimeML 1.2.1 A Formal Specification Language for Events and Temporal Expressions (2005)
20. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. 34(1), 1–47 (2002)
21. Swan, R., Allan, J.: Automatic generation of overview timelines. In: SIGIR 2000: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 49–56. ACM Press, New York (2000)
22. Vila, L.: Handbook of Temporal Reasoning in Artificial Intelligence. In: Formal Theories of Time and Temporal Incidence, Elsevier Science Inc., Amsterdam (2005)
23. Weiss, D.: Stempelator: A Hybrid Stemmer for the Polish Language. Technical Report RA-002/05, Institute of Computing Science, Poznan University of Technology, Poland (2005)

Turning a Corporate Folksonomy into a Lightweight Corporate Ontology

Céline Van Damme[1], Tanguy Coenen[2], and Eddy Vandijck[1]

[1] MOSI
[2] STARlab
Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium
{celine.van.damme,tanguy.coenen,eddy.vandijck}@vub.ac.be

Abstract. Companies use company-specific terminology that may differ from the terminology used in existing corporate ontologies (e.g. Tove) and therefore need their own ontology. However, the current ontology engineering techniques are time-consuming and there exists a conceptual mismatch among developers and users. In contrast, folksonomies or the flat bottom-up taxonomies constituted by web users' tags are rapidly created. In this paper, we present an approach that cost-efficiently derives a lightweight corporate ontology from a corporate folksonomy. We tested it on the folksonomy of a European company and first results are promising: it shows that it creates additional value to the company.

Keywords: ontology, folksonomy, company, added value.

1 Introduction

It has been stated, e.g. in [1,2] that ontologies improve the communication among humans or machines since they provide a shared understanding of a domain. This makes that ontologies are very useful for companies. For instance they can help to improve the communication between employees and to integrate different information systems.

At this moment, there exist several corporate ontologies, for instance Tove [3] and Enterprise ontology [4]. These ontologies describe general concepts and relations related to enterprise and process modeling. We believe these kinds of ontologies may not be useful for every enterprise since companies have a corporate-specific terminology and consequently have their own concepts. In our opinion, an enterprise may need its own corporate ontology.

However, building ontologies with the current ontology engineering techniques have disadvantages. First of all, it is a very time-consuming process[5] and secondly the actual users are not involved in the developing process. As a consequence there exists a conceptual mismatch between the developers and the actual users' vocabulary [6].

These disadvantages are not present in the relatively new categorization method called tagging and its resulting folksonomy. Following the Web2.0 paradigm, a

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 36–47, 2008.

growing number of websites incorporate a tagging/folksonomy mechanism. They allow users to refer to resources (bookmarks, pictures or scholarly publications) on the web with freely selected keywords or tags. The users are not restricted to a controlled vocabulary produced by a group of experts. Users can enter any words that enter their mind. This makes them active participators in creating new tags. Aggregating this user created meta data leads to a flat, bottom-up taxonomy, also known as a folksonomy.

Despite the strengths, tagging has its weaknesses: no conceptual meaning or hierarchical relations are added to the tags. As a consequence, tags have no synonyms or homonyms. Furthermore, specialized as well as general tags can be used to annotate the same resource [7,8]. These weaknesses can be solved by (1)giving the users tools that enable them to add more information to their tags (e.g. cluster tags as on Delicious) [8] and/or (2) trying to generate more information on the tags by employing text mining, statistical techniques and asking additional feedback from the community.

The last few years, we observe a growing attention of the semantic web community for tagging and its resulting folksonomies. At the one hand, we observe researchers that try to enrich the flat ambiguous tags with existing online resources (e.g. Google, Wordnet, existing ontologies) [9] and on the other hand, there are researchers that consider this user created meta data as a valuable source to develop ontologies [10].

In this paper, we argue that cost-efficiently deriving a lightweight ontology from a folksonomy is also applicable to a corporate folksonomy. We regard a lightweight ontology as the simplest form of an ontology: an ontology where only one relation is included or a taxonomy as described by [11]. We propose a 6-step approach which includes several techniques such as the Levenshtein metric, co-occurrence, conditional probability, transitive reduction and visualization. Although, some suggestions have already been made on how a corporate ontology can be built from a corporate folksonomy, no research results have been published so far. We implemented our approach on a corporate folksonomy of a large European distribution company in which Dutch and French are the two official company languages. We obtained the simplest form of an ontology, a lightweight ontology, visualized with the open source tool Graphviz[1]. First results show that our approach is promising and we also detected a number of benefits for the company.

The paper is structured as follows: we provide an overview of related work in section 2. In section 3, we discuss all the techniques and resources suggested in literature to enrich folksonomies or derive ontologies. We explain whether they are suitable to turn an enterprise folksonomy into a lightweight corporate ontology. The most appropriate ones are consequently included in the 6-step approach which is presented in section 4. In section 5, we discuss the general results of our experiment and describe some benefits for the company. Section 6 discusses our findings and presents our future research. A conclusion is provided in section 7.

[1] http://www.graphviz.org/

2 Related Work

At the time of writing, few papers have been written on discussing the use of folksonomies in a company. The authors in [12] present a social bookmarking tool, called Dogear, that lets employees tag their bookmarks from the corporate intranet and the world wide web. The advantages of collaborative tagging in the enterprise is discussed in [13]. The authors suggest that tagging can be used as an expert location tool that facilitates the process of organizing meetings with experts in the company. Tags are a reflection of people's interest and/or knowledge and can as a consequence be seen as a tool to detect experts and their domain of expertise.

However, the authors in [12,13] do not explain how to make the tags less ambigious nor turning them into an ontology. This is discussed in [14]. The authors propose to derive a CRM[2] ontology from a corporate folksonomy. They suggest an integrated visual approach that integrates text mining techniques, tags and user feedback. Each time the employee adds a message or note to the CRM system, tags are required. At the same time, automatic keywords are detected based on the tf-idf score. The tf-idf score is calculated by multiplying the word's document frequency by the logarithm of its inverse document frequency in the set of relevant company documents. The higher the score, the more descriptive the keywords are [15]. In a first phase the user has to indicate whether there exists a relationship between the tags and the keywords with the highest tf-idf score. The relationship has to be specified in a second phase. In this approach, the human effort as well as the implementation time is very high. We also have to point out that the proposed approach still has not been tested.

Literature on folksonomies enrichment or turning folksonomies into ontologies is currently more common in the domain of the World Wide Web. In [16] tags of the photo-sharing site Flickr were used in an experiment to induce a taxonomy, the simplest form of an ontology [11]. The approach of [16] is based on statistical natural language processing techniques where a subsumption or hierarchical relation was deducted. The authors of [9,10] both suggest to include different techniques as well as the wealth of existing online web resources such as Wordnet, Wikipedia, Google, online dictionaries and existing ontologies. The authors in [9] present an approach to enrich tags with semantics to make it possible to integrate folksonomies and the semantic web. The authors use online lexical resources (e.g. Wordnet, Wikipedia, Google) and ontologies to map tags into concepts, properties or instances and determine the relations between mapped tags. However, the resources are tapped in one way (e.g. Wikipedia is used as spelling checker for tags) and the community is not involved to confirm the semantics obtained from existing ontologies and resources. Consequently, tags that reflect new concepts, relations or instances or new relations between tags are neglected. On the contrary, the opposite is suggested in [10]: ontologies are derived from folksonomies. Online lexical resources are suggested to be exploited in several ways. For instance Wikipedia is suggested as a spelling checker as well

[2] CRM or Customer Relationship Management.

as a tool for finding concepts and homonyms. Furthermore, the authors suggest to involve the community.

However, a corporate folksonomy differs from a folksonomy created on the World Wide Web. The users, their underlying motivations and the environment can be different. In case of a corporate folksonomy the user or employee is known and will not always tag voluntarily. An employee may be enforced to tag or may be given an incentive by the company. As a consequence, the amount of additional feedback asked from the users to create a lightweight ontology should be reduced. Labor costs are very high and therefore the number of employees involved with the feedback process should be minimized. In contrast to web communities it is far easier to ask the cooperation of the community: community members have a different mindset than employees and are more willing to participate in additional processes. However, in most cases they are anonymous. Company-specific terminology is mostly used in a closed company environment which makes it hard to include web resources in the ontology construction process. The terminology may contain terms which have a specific meaning for only a small group of employees. In the next section, we will explain how the techniques and resources proposed in literature are applicable to the creation of a lightweight ontology by means of corporate folksonomy.

3 Overview of Techniques and Resources

We can distinguish 2 important steps in the literature to enrich folksonomies or to turn folksonomies into ontologies: (1) finding similar tags and (2) finding concepts and relations between tags. In this section, we briefly describe the different techniques and resources and how they can be applied in each step. We also explain how they can be employed to the creation of a lightweight ontology.

3.1 Step 1: Finding Similar Tags

Stemming algorithms. A stemming algorithm reduce a word to its stem or root. The algorithm removes suffixes and hereby e.g. reduces the words *linked* and *links* to *link* [17]. The algorithm includes rules that are language dependent. Company-specific language can be lost because of the stemming algorithm. These words can differ from the general spelling rules or they can be abbreviations. Some languages, such as Dutch, incorporate English words in the vocabulary without adjustments to the Dutch language. When stemming algorithms are used, there should be a way to determine the language of the tags and whether it involves corporate-specific language.

Levenshtein metric. The Levenshtein metric is a text similarity metric which calculates the distance between two words. More specifically, it counts how many letters have to be replaced, deleted or inserted to transform one word into the other [18]. Dividing this sum by the total number of letters in the word, gives us the Levenshtein metric. It is a valuable technique to verify the similarities of two tags. In order to calculate the distance, first all possible tag pairs have to be made. In [9] a threshold value of 0.83 is used to indicate that two tags are similar. Yet tests showed us that a threshold value of 0.83 excluded a number of similar tags. For instance, the Dutch adjective *groen* or *green* in English, depending on its function in a sentence, can be *groen* or *groene*. Both tags are the same and their Levenshtein similarity is lower than 0.83. We believe this technique should be employed at a lower threshold value and include human feedback, a representative employee that is very well aware of all the terminology used in the company, to confirm or reject the similarity.

Resources. In [9,10] the use of online resources such as Google, Wikipedia, online dictionaries is suggested. These resources can be regarded as spelling checkers. However, the company-specific terminology makes it though hard to use them. For instance, a company had a *gara* tag, used as the abbreviation for the Dutch word *garage*. When using *gara* as a search term for Google, we did not find any link referring to the correct meaning of the term. On Wikipedia, we found a page describing the term, but the concept or description attributed to it was incorrect. On Wikipedia, *gara* is a Basque word and the name of a Spanish newspaper. This causes problems. We have to know whether the tag belongs to the specific terminology of the company or not. In order to find this out, human feedback is necessary.

3.2 Step 2: Finding Concepts and Relations between Tags

Co-occurrence. For each tagged resource all the tag pairs are determined. The tie strength between a tag pair is increased each time two tags are used together. It is interesting to know which tags are often used together in a corporate environment.

Clustering Techniques. These techniques can be used to cluster related tags. In [9] clustering techniques are applied to the co-occurrence matrix of the tags. They calculated the cosine angle between the tag vectors that represent the tags. The smaller the angle, the more similar the vectors are. This can be an interesting technique to find a group of related terms in the company.

Conditional Probability. A rule based on the conditional probability definition was proposed in [16]. More specifically, the rule tries to find out whether one of the tags in the pair can be defined as broader and the other one as narrower term. By applying the definition of the conditional frequency, the conditional probability is calculated by dividing the co-occurrence of the tag pair by the frequency of the individual tag's. Results vary between 0 and 1. The higher the result, the more the term is used in combination with the other term and consequently the more dependend it is of the other term. When the difference between the two results exceeds a certain threshold value, in [16] the threshold value is set to 0.8, a subsumption relationship is found. Finding broader and narrower terms is important to derive hierarchical relations.

Social Network Analysis. These techniques make clusters of people based on their shared objects and or tags. In [10], these techniques are suggested to analyze the ontology of a community. This could be useful to a company, to analyze the ontology of a team.

Transitive Reduction. In [16] the authors remove the roots that are logically above the parent nodes. However, we believe transitive reduction, a technique from graph theory, is far more interesting. Transitive reduction reduces the edges of a graph G to a graph G' by keeping all the paths that exist between the nodes in Graph G [19]. The edges are consequently removed because of the implied transitivity.

Visualization Techniques. The use of visualization is proposed in [14] to lower the barriers to participate in naming the relations between concepts. In literature, several approaches for visualizing tags and lightweight ontologies are described. In [20] CropCircles are suggested to help people understand the complexity of a class hierarchy. We hypothesize that visualizing the lightweight corporate ontology may facilitate the validation process of the approach and might generate added value to the company.

Resources. The resources mentioned in 3.1 as well as the existing ontologies are suggested as a mean to detect concepts and find relations between tags [10]. As mentioned in 3.1, it is hard to use them for a corporate ontology.

4 Deriving a Lightweight Corporate Ontology

In this section, we first describe the corporate folksonomy under study and then proceed to elaborating our approach.

4.1 Folksonomy in the Company

We have implemented our approach in a large European distribution company with headquarters in Belgium in which Dutch and French are the two official company languages. The company employs more than 15.000 people across Europe.

Tagging has been used on all their communication messages for more than 20 years. Messages such as letters and faxes that are not sent electronically are manually scanned, tagged and archived into an information system. Tags replace the subject line of the message. Tagging is completely integrated in the corporate culture. The messages can be created manually, automatically and semi-automatically. The automatic and semi-automatic messages have default tags. In case of semi-automatic messages, the author has to add complementary tags. Manually created messages require user created tags.

Initially, tags were introduced to solve the information retrieval problem since full text search engines were not available at the time. Tagging has remained part of the communication messaging system. However, the ambiguity of the flat tags and the information overload obstructs the search process. The company introduced some tag rules such as a minimum number of tags, no stop words, no plurals and no conjugated verbs, but only a minority of the employees in the company obeys all these rules.

Even though the tagging system at this company is somewhat different from current web-based tagging practices, the 20-years worth of tagged messages represented a real opportunity to test out the approach in a real-life case. Such cases are rare, as not many organizations have adopted tagging in a way which allows the analysis of a large body of tags. Tagging is so widely adopted and part of the corporate culture we believe the tags can be made to represent a non-toy lightweight ontology.

4.2 Statistics of the Corporate Folksonomy

In 2006, more than 7.000.000 messages were created and roughly 60.000.000 tags in total were used. 91% of the messages are created by Dutch speaking employees. We focused our analysis on the tags added to Dutch messages. More specifically, we analyzed 2 different message types individually: quick internal messages and notes since these are the most important communication channels in the company.

4.3 Our Approach

In this paragraph, we present our approach to derive a corporate ontology. The approach includes 6 steps. We included the following techniques: Levenhstein metric, co-occurrence, conditional probability function,transitive reduction and visualization techniques.

Step 1: Selection of the Tags. First, we made a list with all the Dutch stop words[3] and removed the stop words from the database. However, there were not so many stop words in the tags, on average 2% of the tags had to be removed. After filtering the messages with fewer than 2 tags, we had to withdraw the less frequently used tags. We ranked the tags in an absolute frequency table and noticed that the quick internal messages and notes have a group of tags (approximately 150) with a very high frequency (between 5000 and 147.000). We decided it would be interesting to use a top down approach and start to grasp the meaning and interrelations of these frequently used tags.

Step 2: Clean the Tags. Since most of the tag rules, imposed by the organization, are not obeyed, tags are polluted (e.g. plural and singular tags) and need to be cleaned up. We used the Levenhstein similarity metric combined with human feedback. In the Dutch language, there exists some confusion about the spelling of certain words. The letter *c* and *k* are sometimes pronounced in the same way. In the last decade the spelling rules changed several times and as a consequence people do not know whether words should be written with a *c* or *k*. For instance

[3] Based on the list available at
`http://snowball.tartarus.org/algorithms/dutch/stop.txt`

the English word *contract* is sometimes written in Dutch as *contract* or *kontrakt*. Humans, more specifically employees, are able to detect these types of keyword similarity without erasing the company-specific terms.

Based on a trial and error method, we decided to take 0.65 as a threshold value. All the tag pairs that reach a Levenhstein similarity of 0.65 are presented to the user and when two keywords are similar, the user has to check the corresponding check button. Then, the tag with the lowest frequency is replaced with the one with the highest frequency. We decided to implement this rule since we believe that the tag with the highest frequency determines how the word should be written by the wisdom of the crowds in the company [21]. After the adjustment, we reselected the tags in the same way as described in the first step.

Step 3: Co-occurrence. For each message we made all the tag pairs. Then, we counted the frequency of each unique tag pair. The more two tags are used together, the higher this frequency or co-occurrence value. We noticed some tag pairs with a high frequency. Again, we decided to include only the ones with the highest frequency (a frequency of more than 5000) to find the most frequent relations.

Step 4: Finding Broader/Narrower Relations. We wanted to derive the simplest form of an ontology and therefore needed to find the broader/narrower relations between the terms, for instance the relation between *animal* and *dog*. We applied the conditional probability function as described in previous section. Therefore, we divided the co-occurrence of the tag pair by the frequency of the tag itself. We did some manual tests deciding on 0.70 as the most appropriate threshold value. The higher the threshold value, the broader and the less deep the resulting ontology will be. Some parts of the ontology are logically interpretable and we were therefore able to do some logical trial and error trials.

Step 5 & 6: Transitive Reduction and Visualization. Some of the relations are removed by the transitive reduction and they are consequently visualised with the Graphviz tool.

5 First Results

We applied our approach to 2 different message types: quick internal messages and notes. When visually comparing the output of the two message types, we noticed that the 2 generated lightweight ontologies contain different terms. This means that the tag usage between the two message types differs. Consequently, we will need to find a way to map the different partial results into a complete ontology.

We noticed that we have captured other relations than merely broader/narrower or *a kind of* relations. For instance the relation between the tags *name of shop* and *baby*, can not really be considered as a *a kind of* relation but more like a *is related to* relation. It provides more information regarding a stock item of the shop. Therefore, it would be interesting to find a way to capture these different kinds of relations and also check whether we may still apply transitive reduction.

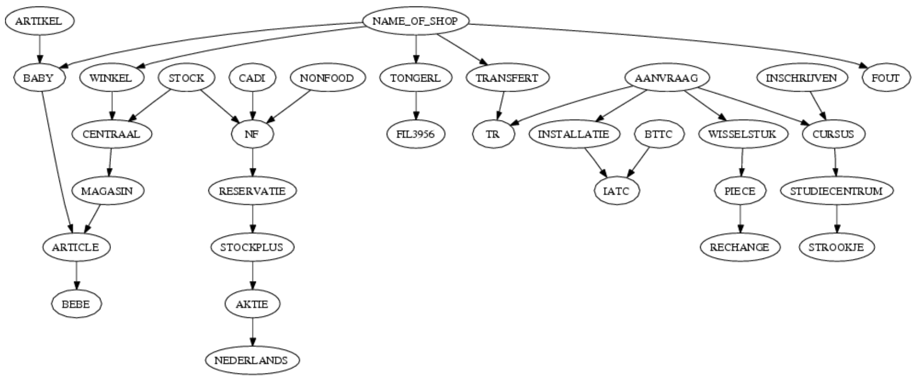


Fig. 1. Part of the lightweight ontology based on tags of quick internal messages

We also observed that the graphs, as in figure 1, include some tags corresponding to the French language such as *article*, *bebe*, *magasin*, *piece*, *rechange*. When having a closer look at the data set, we noticed that there are some bilingual messages with bilingual tags. The tags can not be directly filtered from the database since there is no unique identifier. Looking at the results, we observed a pattern: the same tag relation exists between the Dutch and French tag pair e.g. in figure 1 (artikel, baby) and (article, bebe). We also observed this in the other results which are not visually included in this paper.

Tests with the Levenhstein metric, revealed that we can eliminate some French tags due to the close similarity among both languages e.g. *factuur* in Dutch and *facture* in French. In this way, the Levenshtein metric can reduce the pollution by French tags.

5.1 Added Value for a Company

The tags of the corporate folksonomy under study are assigned to all the communication messages sent in the company. We believe that tags assigned to communication messages reflect the actual business processes in the company. This contrasts with tags used in a social bookmarking tool such as presented in [13]: tags represent the knowledge or interests of the employee. By applying our approach to these tags, we have reduced their tag's weaknesses as described in 1, summarized and visualized all the actions that have taken place in the company over a time period.

Based on the analysis of the visual output generated through our approach, we noticed a number of benefits for the company. As we will explain in the next paragraphs, we believe the visualization obtained from the approach could be used as a tool for management, follow-up tool for new terminology and as a tool for the creation of new teams.

Management Tool. By observing figure 1, we noticed two remarkable relations. On the one hand, we saw that there exists a link between the *name of* *shop*[4] and the tag *fout* or *mistake* in English. On the other hand, we found a relationship between the *name of shop* and the tags *Tongerl* and *Fil3965*. The tag *Tongerl* is used as the abbreviation for a Belgian city and *Fil3965* is the ID of one of the shops. The first mentioned relationship could be a signal that something is wrong and that the relationship between these tags should be further investigated. The latter one could indicate that the shop *Fil3965* has a high sales revenue or high customers complaints. By taking the time factor into account, these results could be compared over different time periods. Therefore, the approach presented in this paper might be an interesting tool for high-level managers in the company. High-level managers are more focused on higher level company's issues such as corporate strategy and are not always aware of all the things that are going on in the company. The visualization of the lightweight ontology obtained through our approach could support them in their daily work and help them in decision making. Therefore, we regard it as a kind of tool for decision making or a sort of add-on for an existing business intelligence tool. This technique could be a valuable extension to existing datamining techniques. At this moment the company is not doing any analysis of its unstructured information.

Follow-up Tool for new Terminology. The proposed approach could be valuable as a follow-up tool for new corporate terminology. It reveals how new terms are utilized and interpreted. In case of company acquisition, such a tool could be very interesting. When a company gets acquired by another company, the acquired company will have to apply new terminology to improve the communication process between both of them. Again, the time factor can be included in the process to evaluate and compare the results.

Creating Teams. When new teams have to be set up, the approach might helpful to choose the most appropriate employees. This visualization shows how tags are combined with other ones. By selecting all the terms that are related to a concept, the corresponding employees could be selected for the creation of a new team.

6 Discussion and Future Research

The approach is briefly validated by presenting the results to the IT-director and the communication system's analyst of the company. They verified the results by looking at the visualizations and checking the tags in the communication messaging system. They both confirmed their validity.

We also discussed the added value of the visualization. In their opinion, the first and third benefit would be most interesting to their company. They even suggested a visual search tool as an additional added value. Such as tool could be an extension of the suggested management tool. When the manager finds an interesting hierarchical relation or cluster, he should be able to click on it to retrieve the corresponding messages.

[4] We renamed this tag to guarantee the anonymity of the company.

We plan to expand our tests to other message types and include tags with a lower frequency to verify the added values we deduced from our current results. In addition, we should set up focus groups with employees of the company where the added value can be extensively discussed.

Further, we will try to find a method to map the ontologies obtained by applying the approach to different message types.

The approach should be further extended and include more techniques and algorithms such as clustering techniques. In this way, more relations might be included in the ontology. However, we believe a cost-benefit analysis should also be built-in in the approach to evaluate whether a more extended version of the ontology will generate the necessarily return on investment. Currently, the approach minimizes the human input and in this way a lightweight-ontology is cost-efficiently derived from the corporate folksonomy.

7 Conclusion

In this paper we proposed a 6-step approach to turn a corporate folksonomy into a corporate taxonomy. We implemented the approach on a folksonomy of a European distribution company where tags are assigned to communication messages. We applied the approach to two different kinds of communication messages: quick internal messages and notes. Based on the results, we concluded that the approach summarizes all the actions that have taken place in the company and reflect the actual business processes. The visualization of the results facilitated the validation process and allowed us to detect some benefits for the company: management tool for managers, a follow-up tool for new terminology and a tool which supports the creation of teams. The first and third mentioned benefits were validated by the IT-director and the communication system's analyst of the company. They even suggested to extent the management tool with information retrieval functionality.

Though the approach should be extended and further tested in the near future, we may conclude that the first results look promising.

References

1. Uschold, M., Gruninger, M.: Ontologies: Principles, Methods, and Applications. Knowledge Engineering Review 11(2), 93–155 (1996)
2. Fensel, D.: Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. Springer, Heidelberg (2003)
3. Fox, M.: The Tove Project towards a common-sense Model of the Enterprise. In: Belli, F., Radermacher, F.J. (eds.) IEA/AIE 1992. LNCS, vol. 604, pp. 25–34. Springer, Heidelberg (1992)
4. Uschold, M., King, M., Moralee, S., Zorgios, Y.: The Enterprise Ontology. Knowledge Engineering Review 13, 31–89 (1998)
5. Paslaru Bontas, E., Tempich, C.: Ontology Engineering: a Reality Check. In: Meersman, R., Tari, Z. (eds.) OTM 2006. LNCS, vol. 4275, pp. 836–854. Springer, Heidelberg (2006)

6. Hepp, M.: How Reality Constraints the Reality of the Development of Relevant Ontologies. IEEE Internet Computing 11(1), 96–102 (2007)
7. Golder, S., Huberman, B.A.: Usage of Patterns of Collaborative Tagging Systems. Journal of Information Science 32(2), 198–2008 (2006)
8. Guy, M., Tonkin, E.: Tidying up Tags? D-Lib Magazine 12(1), 1082–9873 (2006)
9. Specia, L., Motta, E.: Integrating Folksonomies with the Semantic Web. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 624–639. Springer, Heidelberg (2007)
10. Van Damme, C., Hepp, M., Siorpaes, K.: Folksontology: An Integrated Approach for Turning Folksonomies into Ontologies. In: Proceedings of Bridging the Gap between Semantic Web and Web 2.0, Innsbruck, pp. 71–84 (2007)
11. Uschold, M., Japsper, R.: A Framework for Understanding and Classifying Ontology Applications. In: Proceedings of the IJCAI 1999 Workshop on Ontologies and Problem-Solving Methods, Stockholm (1999)
12. Millen, D.R., Feiberg, J., Kerr, B.: Dogear: Social Bookmarking in the Enterprise. In: Proceedings of Collaborative Web Tagging Workshop at WWW 2006, Edinburgh (2006)
13. John, A., Seligmann, D.: Collaborative Tagging and Expertise in the Enterprise. In: Proceedings of Collaborative Web Tagging Workshop at WWW 2006, Edinburgh (2006)
14. Van Damme, C., Christiaens, S., Vandijck, E.: Building an Employee-driven CRM Ontology. In: Proceedings of the IADIS Multi Conference on Computer Science and Information Systems (MCCSIS): E-society 2007, Lisbon, pp. 330–334 (2007)
15. Salton, G., McGill, J.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (1986)
16. Schmitz, P.: Inducing Ontology from Flickr Tags. In: Proceedings of Collaborative Web Tagging Workshop at WWW 2006, Edinburgh (2006)
17. Porter, M.F.: An Algorithm for Suffix Stripping. Program 14(13), 130–137 (1980)
18. Levenshtein, V.I.: Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. Soviet Physics Doklady 10(8), 707–710 (1966)
19. Aho, A.V., Garey, M.R., Ullman, J.D.: The Transitive Reduction of a Directed Graph. SIAM J. Comput. 1(2), 131–137 (1972)
20. Parsia, B., Wang, T., Goldbeck, J.: Visualizing Web Ontologies with CropCircles. In: Proceedings of the End User Semantic Web Interaction Workshop at ISWC, Galway (2005)
21. Surowiecki, J.: The Wisdom of the Crowds. Anchor, New York (2005)

Discovering Semantics in Multimedia Content Using Wikipedia

Angela Fogarolli and Marco Ronchetti

University of Trento,
Dept. of Information and Communication Tech.,
Via Sommarive 14, 38050 Trento, Italy
{angela.fogarolli,marco.ronchetti}@unitn.it

Abstract. Semantic-based information retrieval is an area of ongoing work. In this paper we present a solution for giving semantic support to multimedia content information retrieval in an e-Learning environment where very often a large number of multimedia objects and information sources are used in combination. Semantic support is given through intelligent use of Wikipedia in combination with statistical Information Extraction techniques.

Keywords: Content retrieval and filtering: search over semi-structural Web sources, Multimedia, Wikipedia, e-Learning.

1 Introduction

Nowadays, organizations have to deal with information overloading. They need a way to organize and store their content and being able to easily retrieve it when necessary. Our objective is to provide a system for indexing and retrieving content based on the semantic provide by Wikipedia. Retrieving the desired content can be difficult due to the the high specifically of terms in a search task.

In our work, we are addressing the problem of accessing different kinds of unstructured or semi-structured information sources taking advantages of the semantic provided by public available resources such as Wikipedia. Furthermore using the approach we will describe in section 3 we would like to automatize the task of annotating a corpus and discover relations between annotations. Next we will use annotation in combination with textual information retrieval for determining the search context and based on it we will be able to give search suggestions and perform query expansion. Using annotation in information retrieval is not a new idea [6, 4] even in combination with ontologies [3], it has been widely used in video and image retrieval generating also a social phenomena like folksonomy [14, 12]. What is new is the use of domain independent public available semantic to automatically describe content in different kind of media.

We are applying our approach in the e-Learning context, specifically enhanced streaming video lectures (see [8, 5]) because of the peculiarity in this scenario of combining different kinds of unstructured or semi-structured sources of information. E-Learning presents many problematics in common with the business

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 48–57, 2008.

scenario in terms of content classification for its amount of information to classify and for the different contexts where a specific information can be relevant. Our target repository collects different kinds of media (video, audio, presentation slides, text documents), which can be searched and presented in combination. For each recorded event [1] we provide not only the video but also related materials, which can consist of presentation slides, documents or Web sites the speaker points to. All the resources are temporally synchronized with the video.

We can summarize the following five state of the art approaches to multimedia indexing and navigation:

1. Use of metadata to browse keyframes.
2. Use text from speech, using transcript-based search.
3. Matching keyframes vs. querying of images. Keyframes extracted as shot representatives are used for retrieval. It requires user to locate images/other keyframes, from browsing or other search.
4. Use of semantic features. They are based upon pre-processing video or keyframes to detect features. Features can be related to ontologies.
5. Use video/image objects as queries.

We concentrate on pt. (2) and partially on pt. (3), we use the text-from-speech technique combined with a textual analysis of the speech and the event related material using Wikipedia instead of ontologies.

The combination of information extracted from video and related material gives a complete picture of an event, since in the real world the sum of *all* the media used by a speaker is meant to fully describe the event's topics to facilitate knowledge transfer to the audience.

In this paper we report about how we provide semantic support and unsupervised annotation of multimedia material based on information extracted from Wikipedia, rather than the usage of Semantic Web technologies (specifically without ontologies). Our approach is domain independent, and in theory it could also be applied to different use cases where there is a need for clustering or annotation of a corpus.

The structure of this paper is organized as follows: in the next section we describe the context and the motivation of our work; section 3 gives an overview of our approach. In Sect. 4 we apply the approach described in the previous section to our use case. Finally, we discuss the directions we are planning to take regarding further work.

2 Semantics in the Web

In the Web, some collections of data containing semantic annotations [2] are now available and there is a trend to semantically enable more and more Web content. Even though this trend is perceivable, there is still a huge amount of material

---

[1] E.g: lecture, seminar, talk, meeting.

[2] UniProt `http://www.ebi.ac.uk/swissprot/index.html`,
Ecademy: `http://www.ecademy.com`

on which these technologies have not been applied. One limiting factor for a faster adoption of Semantic Web technologies, is the difficulty to find ready-to-use conceptualizations for annotating existing material and making it Semantic Web compatible.

We explored the possibility of using Semantic Web ontologies for annotating multimedia material and for discovering and presenting to the user relations between the searched topics and other topics, based on the relationships between entities in one or more domain ontologies. We experienced difficulties in finding ontologies which cover a variety of domains, since e-Lectures can cover an unpredictable amount of domains (e.g. computer science, history, meteorology, geography, math...). In addition, the terms expressed in e-Lectures are usually *individuals* of an ontology [3] and finding populated ontologies with a wide coverage of individuals to date is a big challenge, and usually requires the involvement or a knowledge engineer.

Our requirement was to find a broad, domain-independent collection of individual terms (as opposed to concepts) which are connected by relations. To the best of our knowledge, the most complete collection of this kind is Wikipedia [4]. Wikipedia contains a classification of topics, organized with an hierarchy of categories and with relationships between elements. The advantage of using it is that the social collaborative network around it makes its content always up to date and it covers in details a huge amount of topics in different domains and languages. In addition it also takes into account the different possible meanings of a term through a disambiguation page.

What we can extract using Wikipedia are the relationships between topics. According to Obrst's definitions in [10], Wikipedia not only offers *weak semantic* information, such as parent-child relationships, but it also contains lexicographic relationships that – once the domain of interest is determined – can offer *medium semantic*. In Wikipedia we do not have *strong semantic*, i.e. we can not describe real-world relationships such as "a car has a minimum of four wheels" as with the usage of an ontology. We can only deduce that concepts are connected without knowing how; we can tell that one concept in one category is related to other concepts which are linked in the description of the concept itself.

In Wikipedia, the concept of class and instance are not separated as in the ontological sense, due to the fact that it is not constrained to a formal model, for the reason of which it is not possible to formalize reasoning on the Wikipedia content directly. There are projects (see section 5) that try to embed semantic inside Wikipedia extending the Wiki software used to write Wikipedia pages [15], and some others [5] which provide an RDF representation of Wikipedia, to make its content machine-interpretable.

[3] E.g. the term 'Collection' in a Java Programming class could be modelled as an instance of a data container class in a Java Programming ontology.

[4] Wikipedia is a freely available encyclopedia which is constantly growing in size and in fame thanks to the copyleft license that allows the content to be copied, modified and redistributed as long as there is an acknowledgment of the author and the new content is published under the same license. See `http://www.wikipedia.org`

[5] E.g. `www.dbpedia.org` [2].

We use Wikipedia as a taxonomy to obtain lexicographic relationships and in combination with statistical information extraction we can deduce related concepts to the terms extracted from our corpus. In addition, since our corpus covers a representation of a part of the real world we also use the corpus itself as "training data" for domain disambiguation in Wikipedia.

3 A Semantic Discovery Approach

In order to discover the semantic present in a corpus we first have to extract and identify the terms from it. Once we have the list of the words contained in each unit of the collection, we can link them through the relationships contained in Wikipedia. We divide the explanation of this approach in two parts: the first one deals with how Information Extraction has been done and the second describes how through Wikipedia one can find semantic relationships. The two steps are independent in the sense that Information Extraction can be carried out in different ways while our Wikipedia module could still be used to find relations between terms. We give here an explanation of the first part only for the sake of contextualization.

3.1 Information Extraction

We performed Information Extraction (IE) by using Lucene [6], a state of the art tool which provides Java-based indexing and search technology using a statistical approach. Lucene had been used in the project as a search engine for querying an unstructured e-Learning repository, but since it also provides basic APIs for analyzing text, we exploited Lucene also for extracting information from our corpus. In general term extraction tools using a statistical approach basically look for repeated sequences of lexical items. We also explored a linguistic approach based on Natural Language Processing (NLP) using other state of the art tools in the area such as GATE [7] and IBM UIMA [8], but the approach was not suited for our use case since it is language and grammar dependent. In fact, in e-Learning the material can be in different languages, and sometimes more than one language can be combined in the same event. For instance, in some cases presentation slides are written in English while the speech is delivered in another language (e.g. in Italian). As a consequence, the work needed to adapt a linguistic approach to our needs was excessive. Moreover, story telling does not play an important role in e-Learning - at least not in the disciplines we considered - and this makes it difficult to locate and classify atomic elements in text into predefined categories for Entity Recognition. For these reasons we chose a statistical approach and we calculated a term vector for each document in our index. The term vector contains a list of terms with their frequency in a document.

[6] http://lucene.apache.org/
[7] http://gate.ac.uk/
[8] www.research.ibm.com/UIMA/ [17].

In order to calculate the term vector we had to store our multimedia material into an index. Many documents can refer to the same event. For instance, we have at least two main information sources for each event: presentation slides and video transcript. Following our multimodal view, we modelled all the event's multimedia material into a single document in the index. In this way the term vector calculated through Lucene is factual. For improving the performance of this task we are currently working on the indexing phrase and in particular in the pre-processing task [9].

3.2 Semantic Extraction Using Wikipedia

In this section we explain how we enhance information retrieval based on the recognition of the relations between topics. Usually, e-Learning users (typically students) do not have a rich understanding of the domain or of how one topic is connected to others. For this reason a tool, which has the goal of enabling access to information, must guide users to retrieve exactly what they are searching for.

For example, a user looking for the term 'Collection' in a Java programming class must find out about the different types of Collection such as 'HashMap', 'Map' and 'Set' since these terms can also be found in the lecture material, and they all mean Collection. Understanding relationships between terms in our corpus permits also us to automatically discover the important topics of an event which can be used for unsupervised classification of the material.

Our starting point for the second phase is the list of terms which were extracted from multimedia material(video, slides, documents) during the first phase. In Wikipedia, we look up the most important extracted terms from the corpus. The goal of this phase is to find a Wikipedia definition page for every important term and to try to extract relations to other terms by examining the hypertextual links in the page. This is done by processing the links in the page. Therefore, the term of interest is found in Wikipedia, and all the links in its page are analyzed.

The process can be divided in four steps:

1. *Finding all the links for a page*
   For each extracted term we search for pages in Wikipedia which contain the term in their name. In Wikipedia every page is named by a string composed of topic name and topic domain. After that, we collect the links for every page. The search is made on a local copy of the English version of the Wikipedia database, but we could also reach the same result by downloading and parsing Wikipedia Web pages. We chose to maintain a copy of the database to increase the speed of the task.
2. *Domain Disambiguation*
   The second step deals with domain disambiguation. A word can have multiple Wikipedia definitions because it can assume different meanings (senses) in different domains. Among the (possibly) multiple Wikipedia definitions,

[9] E.g. cleaning the text from Italian or English stopwords and applying different language stemmers as a filter, building categorizer for improving the quality of the raking of the extracted lecture terms.

we choose the one which has the most link words in common with the extracted lecture's terms. We manually checked this approach to evaluate the accordance of the semantic expressed in the disambiguated terms with the one of the event and we find out that this is true for the majority of the cases.

3. *Individuation of the 'strong' links*
   We define a link to be 'strong' if the page it points to has a link back to the starting page. For instance, "Rome" and "Italy" are strongly linked since the page on Rome says that it is the capital of Italy, while the page on Italy reports that Rome is the capital of the state. A minor town located in Italy will instead have a 'weak' link with Italy, since in its page it will be declared that the town is in Italy but in the page for Italy the minor town will most likely not be mentioned. In our case, strong links are candidates for topics related to the searched term, and they will be used for giving user suggestions in query expansion and in the process of summary generation of Wikipedia definitions.
4. *Annotation through Wikipedia definition summarization*
   In this last step we use the extracted strong links for every important word of an event to automatically generate a summary of the word definition in Wikipedia.The summary is generated taking all the sentences from the Wikipedia definition page in which a strong link is present; usually fifty percent of the content of original definition is selected. The summary is then used for expressing the meaning of the important term. In other words, we annotate the lecture through Wikipedia terminology, and for each term we keep a brief definition.

An example follows to clarify the process. Suppose one of the terms extracted from the e-Lecture material is 'Collection', which is in the list of the extracted keywords. Consider a simplified list of extracted keywords (in multiple languages!) as follows: *Elemento, Map, Tipo, Object, Method, Interface, computer science, Collection, Oggetto.*

1. The first step of the algorithm will search for all the pages which contains the term *'Collection'* in its name. A search in the Wikipedia database will find a relatively large number of pages that satisfy this requirement due to the different meaning the word can have. Consequently we retrieve the links of every found page. We will use the links in the disambiguation step.
2. The second step calculates domain disambiguation.
   We automatically identify the right Wikipedia entry based on the domain defined by the multimedia material the keyword was extracted from. So we select one page among the ones we retrieved during step one which is in concordance with the domain of interest. The disambiguation function considers for every candidate page its links. In particular it looks for correspondences among the link's names and the keywords extracted from the corpus. The page which has the largest number of links in correspondence with the corpus' terms will be considered to be the correct one and it will be used as the disambiguated term. The term that has been disambiguated has the

same meaning in Wikipedia and in the corpus. The result of this step is the identification of the disambiguated terms with their links.

3. In the third step we identify the strong links between the ones we found through the disambiguation step. We will use the strong links in our application for querying support and for automatic summarization of the content of Wikipedia entries.
4. In this last step we use the extracted strong links for every important word of the lecture to automatically generate a summary of the word definition in Wikipedia. We download and parse every Wikipedia page which refers to the most important words in our event. We extract from the Wikipedia definition all the sentences which contain a strong link and the term itself. Putting all the extracted sentences together permits to generate a reasonably well written summarization of the term definition.

4 Applications

In this section we describe some applications of our approach we implemented in our use case application. Many other applications are under considerations. We use the approach for the following functionalities:

- *Search Suggestion and Query expansion*
  Wikipedia is used for finding topics related to the searched one. In our search user interface we show the hits for the searched string and a bunch of links to some related topics which have a correspondence in our repository. A click on one of the link will initiate a search for the occurrences of the link term in the learning material. This is done by viewing all the strong links retrieved through Wikipedia which term appear also in the event material, in this way we can suggest different search terms or topics that are connected to the first searched one.
- Automatic annotation
  For each occurrence displayed in the hits, we show some links to related important topics. The important topics automatically annotate the event with some terms which have a predefined meaning in Wikipedia. In this way there are no more ambiguities in the meaning of a term used for annotation. Another advantage of the strong link identification in combination with the term vector extracted for every event is the possibility to automatically describe the most important concepts of the event.
- Automatic summarization
  The semantic discovery approach described in the previous section brought us to the individuation of the strong links for each topic. Based on them we can generate for each event annotation (topic) a brief summarization of the description of the topic in Wikipedia. A click on one of the event annotation will display the summary plus the retrieved hits for that term. In our search user interface for each event(lecture, seminar, meeting) we show the six most important words and the related summarized Wikipedia definitions.

5 Related Work

Wikipedia contains a vast amount of information, therefore there have been mainly two approaches for exploring its content and make it machine readable. The first approach consists in embedding semantic notations in its content [15, 7]; while the other one deals with information extraction based on the understanding of how the Wikipedia content is structured: [1, 13, 16, 11, 18].

The SemanticWikipedia project [15] is an initiative that invites Wikipedia authors to add semantic tags to their articles in order to make them machine interpretable. The wiki software behind Wikipedia(MediaWiki [7]), itself enables authors to represent structured information in an attribute-value notation, which is rendered inside a wiki page by means of an associated template.

The second main stream of Wikipedia related work is on automatically extract knowledge from the Wikipedia content as in [1, 13, 16, 11, 18].

DBpedia [1]is a community effort to extract structured information from Wikipedia and to make this information available on the Web. DBpedia offers sophisticated queries against Wikipedia and to other linked datasets on the Web. The DBpedia dataset describes 1,950,000 "things", including at least 80,000 persons, 70,000 places, 35,000 music albums, 12,000 films. It contains 657,000 links to images, 1,600,000 links to relevant external web pages and 440,000 external links into other RDF datasets. Altogether, the DBpedia dataset consists of around 103 million RDF triples. DBpedia extracts [2] RDF triples from Wikipedia informations presented in the page templates such as infoboxes and hyperlinks.

Yago [13] is a knowledge base which extends the relationships of DBpedia extending the standard RDF notation. At December 2007, Yago contained over 1.7 million entities (like persons, organizations, cities, etc.) A YAGO-query consists of multiple lines (conditions). Each line contains one entity, a relation and another entity.

DBpedia or Yago could replaced Wikipedia as a source of knowledge in our semantic discovery approach, although at the time of this writing these knowledge bases contain only entities (such as person and places) and not abstract concepts as the one we have in e-Learning material. In addition we don't know a priori with which properties a term a can be searched, so in our domain replacing Wikipidia free-text would not be beneficial.

ISOLDE [16] is a system for deriving a domain ontologies using named-entity tagger on a corpus and combining the extracted information with Wikipedia and Wiktionary. The results shows that this kind of approach works better with semi-structure information such as dictionaries.

KYLIN [18] is another project which aim is automatically complete the information presented in the Wikipedia infoboxes analyzing disambiguated text and links in Wikipedia pages.

Ponzetto et al. [11] in their work have explored information extraction on Wikipedia for creating a taxonomy containing a large amount of subsumptions, i.e. is-a relations.

6 Future Directions

One of the activities we plan for the near future is related to extending the Wikipedia module to support various languages. The multilanguage support consist in recognizing relations between terms in the corpus which are not in English. As a first step, we'll look at the links to the other instances of Wikipedia in different languages. In most cases, pages in the Wikipedia instance in one language have links to pages in many other Wikipedia instances in other languages. Since these links were created manually by the page authors, in most cases there is no ambiguity in the translation. In case a link to the target language of interest is not present, we can resort to freely available, albeit less trustable external sources for translating from and to English. The Wikipedia process described in the previous section will not change but writing language dependent processing modules such as language specific stemmers should be added to enable the comparison of the related Wikipedia content found in English with the terms contained in the multimedia content repository. Consequently we have scheduled an evaluation of the presented approach for annotating a large amount of text resources and a user based evaluation to assess if the introduction of semantic multimedia information retrieval is actually bringing an advantage to the student. We will carry out a student's performance evaluation on some topics presented in the e-Learning repository and we will compare the results with the the ones we gathered last year using a text based search that was not semantically enhanced [9].

7 Conclusion

In this paper we described an approach to semantically annotate the content of an unstructured multimedia repository. The annotation has been done combining the terms extracted from the corpus with lexicographic relationships from Wikipedia. Wikipedia has been used as an alternative to ontologies. The content annotated in this way permits to keep track of the relations between annotations. The approach has been used for giving search suggestions in multimedia information retrieval, in multimedia annotation and for giving a brief description of the topics of the multimedia event. Our approach is domain independent, and it could in theory also be applied to different use cases where there is a need for clustering or annotation of a corpus.

References

[1] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.G.: DBpedia: A Nucleus for a Web of Open Data. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ISWC 2007. LNCS, vol. 4825, Springer, Heidelberg (2007)

[2] Auer, S., Lehmann, J.: What have innsbruck and leipzig in common? extracting semantics from wiki content. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, Springer, Heidelberg (2007)

[3] Bertini, M., Bimbo, A.D., Torniai, C.: Enhanced ontologies for video annotation and retrieval. In: MIR 2005: Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval, pp. 89–96. ACM, New York (2005)
[4] Bontcheva, K., Maynard, D., Cunningham, H., Saggion, H.: Using human language technology for automatic annotation and indexing of digital library content. In: Agosti, M., Thanos, C. (eds.) ECDL 2002. LNCS, vol. 2458, Springer, Heidelberg (2002)
[5] Dolzani, M., Ronchetti, M.: Video streaming over the internet to support learning: the lode system. WIT Transactions on Informatics and Communication Technologies 34, 61–65 (2005)
[6] Dowman, M., Tablan, V., Cunningham, H., Popov, B.: Web-assisted annotation, semantic indexing and search of television and radio news. In: Proceedings of the 14th International World Wide Web Conference, Chiba, Japan (2005)
[7] Ebersbach, A., Glaser, M., Heigl, R.: Wiki: Web Collaboration, November 2005. Springer, Heidelberg (2005)
[8] Fogarolli, A., Riccardi, G., Ronchetti, M.: Searching information in a collection of video-lectures. In: Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2007, Vancouver, Canada, June 2007, pp. 1450–1459. AACE (2007)
[9] Fogarolli, A., Ronchetti, M.: Case study: Evaluation of a tool for searching inside a collection of multimodal e-lectures. In: Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2007, Vancouver, Canada, June 2007, pp. 3893–3900. AACE (2007)
[10] Obrst, L.: Ontologies for semantically interoperable systems. In: CIKM 20003: Proceedings of the twelfth international conference on Information and knowledge management, pp. 366–369. ACM Press, New York (2003)
[11] Ponzetto, S.P., Strube, M.: Deriving a large scale taxonomy from wikipedia. In: Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI 2007), Vancouver, B.C (2007)
[12] Porter, J.: Folksonomies: A user-driven approach to organizing content. User Interface Engineering (2005)
[13] Suchanek, F., Kasneci, G., Weikum, G.: Yago: A large ontology from wikipedia and wordnet. Research Report MPI-I-2007-5-003, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany (2007)
[14] Tennis, J.T.: Social tagging and the next steps for indexing. In: 17th SIG/CR Classification Research Workshop (2006)
[15] Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., Studer, R.: Semantic wikipedia. In: Proceedings of the 15th international conference on World Wide Web, WWW 2006, Edinburgh, Scotland, May 23-26, 2006 (2006)
[16] Weber, N., Buitelaar, P.: Web-based ontology learning with isolde. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, Springer, Heidelberg (2006)
[17] Welty, C., Murdock, J.: Towards knowledge acquisition from information extraction. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 709–722. Springer, Heidelberg (2006)
[18] Wu, F., Weld, D.: Autonomously semantifying wikipedia. In: ACM Sixteenth Conference on Information and Knowledge Management (CIKM 2007), Lisbon, Portugal (November 2007)

Combining Structure and Semantics for Ontology-Based Corporate Wikis

Alexandre Passant[1,2] and Philippe Laublet[1]

[1] LaLIC, Université Paris IV,
28 rue Serpente,
75006 Paris, France
firstname.lastname@paris4.sorbonne.fr
[2] Electricité de France Recherche et Développement,
1 avenue du Géneral de Gaulle,
92141 Clamart Cedex, France
firstname.lastname@edf.fr

Abstract. While wikis offer new means to collaboratively build, organize and share knowledge in organizations, such knowledge cannot be easily understood by computers in a query answering or reusability process. This paper details the features and architecture of a wiki-farm system that combines structure and semantics in order to collaboratively produce ontology-based data and immediately reuse it in wiki pages to enrich browsing and querying capabilities of the system.

Keywords: wikis, ontologies, Semantic Web, Web 2.0, architecture of participation, collective intelligence.

1 Introduction

During the last few years, different services and principles of Web 2.0 appeared in enterprise environments, such as blogs, wikis, RSS feeds or tagging [1]. On the other hand, Semantic Web [2] technologies are used in different business information systems to enrich data integration, querying and browsing, thanks to powerful means to represent knowledge like ontologies. While some think that Web 2.0 and Semantic Web are opposite concepts, we believe, as others [3], that these two views about the future of the Web should - and even must - be combined to offer easy-to-use but powerful services to end-users. Especially, in a business context, we think that information systems should benefit from usability and social aspects of Web 2.0 and also from data formalisms of the Semantic Web to provide to end users means to collaboratively build, maintain and re-use ontology-based data, a task often dedicated to knowledge management experts.

This paper, which introduces a semantic wiki-farm prototype that aims to achieve this goal, is organized as follow. First, we will briefly detail how wikis can be used in business information systems and the reasons why they are limited for reusing data. We will then introduce how Semantic Web and ontologies can be used to enhance them, through a state of the art of semantic wikis. Then, we will

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 58–69, 2008.

explain in detail our views regarding a wiki-farm system that combines structure and semantics to enhance knowledge management within organizations. We will detail its different components and the way users can use it to collaboratively build and reused formalized data. Finally, we will conclude with some statistics about the use of the system at EDF R&D[1] and will end the paper with some perspectives and future works regarding the software.

2 From Wikis to Semantics

2.1 Wikis for Knowledge Management in Organizations

Among the numerous practices and tools that became popular within Web 2.0, such as tagging, RSS feeds, blogs and wikis, the latter offers new and interesting possibilities regarding collaborative knowledge management in organizations since ideally, anyone could add information to a given field.

Yet, bridging this gap between traditional information systems - with access rights and workflow management - and such ways to manage knowledge implies social and organizational changes that are not always easy to accept for end-users, in environments when, most of the time, knowledge equals power. At EDF, we setup several wikis for different communities of practices and while some of them first decided to completely open their platform, other groups restricted it in reading, editing or both, by fear to see their knowledge spread in the company without any reward or feedback that could enrich their own. We noticed that it mostly depends on users' background, since communities dealing with topics as nuclear energy were more into a close community approach, while those involved in open-source software opened their platforms and even blogged about it. Nevetherless, two wiki features made some users change their mind: (1) when closed communities realized that open ones got feedback from their wiki, some of them opened it and (2) history pages let users feel rewarded for their work since they can still claim a kind of knowledge ownership on open wiki pages.

Nevertheless more than this social aspect, wikis suffer from the difficulty for computers to exploit the knowledge they contain. Indeed, a reader could learn from a wiki that EDF is based in France and produces nuclear energy, but a software agent will not be able to easily answer queries like *"Is EDF located in France ?"* or *"List all companies known in that wiki"*. The reason is that wikis deal with documents and not with understandable representations of real-world objects, as a human reader does. Thus, a wiki will model that *"There are some hyper-links between a page titled EDF, a page titled France and a page titled nuclear energy"*, but will not be able to deduce anything about the nature of those different objects and the relationships between them, since pages do not carry enough semantics about what they contain (Fig. 1).

[1] Electricité de France (EDF) is the leading energy company in France and has 3 research sites, employing about 2000 people. See http://rd.edf.com

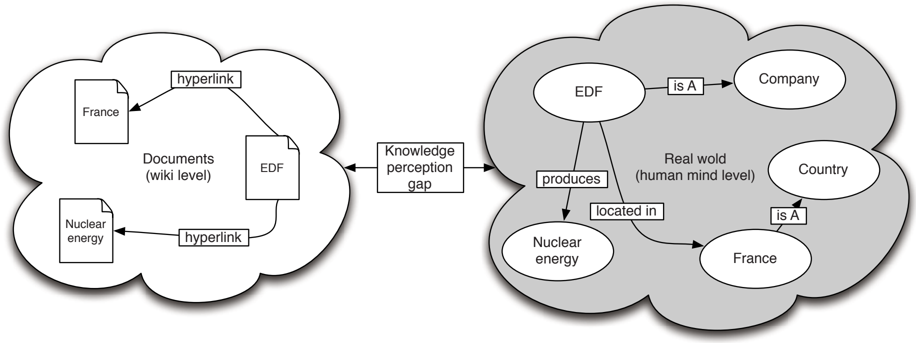


Fig. 1. The gap between documents and real-world knowledge

2.2 The Semantic Web and Ontologies for Better Wikis

To bridge this gap between documents and machine-readable knowledge, data must be described in a way that software agents interpret and understand in a uniform manner. Ontologies [4], as defined in the Knowledge Engineering field, are an effective way to do so, since they provide common vocabularies and structures for modeling domains of interest and thus can be used by different applications of information systems [5]. Moreover, the Semantic Web initiative, which goal is to provide a common framework for such data modeling on the Web, envisions a Web of Data [6] or Global Giant Graph [7]. It defined standards and languages to achieve those tasks, such as RDF to represent data in an uniform way, RDFS or OWL to model ontologies and SPARQL [8], which recently became a W3C recommendation, to query RDF data. Using ontologies and Semantic Web technologies in business information systems can help to achieve various tasks [9] such as advanced knowledge representation, integration of heterogeneous data or even reasoning over existing datasets.

Thus, during the last few years, various semantic wikis prototypes have been built, combining wiki features and Semantic Web technologies. Systems such as Semantic MediaWiki [10] or SemperWiki [11], require to use a special wiki syntax or to embed RDF, to add semantic annotations in wiki pages. While it is an open approach, it can lead to semantic heterogeneity problems since any user can use its own vocabulary to add annotations in a document, making those annotations difficult to re-use. A system like IkeWiki [12] combines plain-text feature of wikis and a dedicated interface to help users annotating content, re-using existing ontologies, while OntoWiki [13] can be used as a complete ontology editor, with a user-friendly interface that offers different views over existing data. Yet most of those systems require users to have some knowledge of Semantic Web, since they have to deal with namespaces or URIs, which make the tools difficult to integrate in business environments where people are not necessary computing experts.

In these tools, created annotations are used to create ontology instances and relationships between those instances, as well as properties, thus providing a real-world and machine-readable representation of the content which is described inside the pages. They can help to enhance browsing capabilities of the wiki, by suggesting related pages sharing similar instances or listing all pages featuring a certain property as does Semantic MediaWiki. Moreover, new ways to browse the data are available in some systems as OntoWiki with its map and calendar view, while some of them feature a back-end RDF store to query data from the whole wiki. Finally, some tools also feature inferencing capabilities in order to deduce new knowledge from the current state of the wiki and thus enrich user experience in discovering new knowledge. For example, IkeWIki and OntoWiki, can list all instances of a given type taking into account instances of various subclasses.

Finally, it seems important to reference DBpedia [15], a project that aims to represent the whole wikipedia content in RDF, as well as other semantic wikis, like SweetWiki [14] which does not focus on ontology population but on using semantic web technologies to let users tag their pages and organize those tags.

3 A Semantic Wiki-Farm for Corporate Knowledge Management

3.1 Main Features

Regarding various aspects of semantic wikis that have been mentioned before and according what seems important to us in a business environment, we designed a new wiki farm - i.e. a wiki server for several different wikis - system, based on the following features, that will be described in the rest of the paper:

- *Ontology-based knowledge representation.* Data created from wiki pages is represented in RDF and is based on a set of ontologies defined by administrators of the wiki in order to avoid semantic heterogeneity problems;
- *Usability.* In extent of the previous point and in order to let users easily produce that ontology-based data, we focused on a combination of plain-text and user-friendly form-based approach to edit wiki pages, so that users do not confront to a new syntax and simply use forms to create RDF data;
- *Combination of data and meta-data modeling.* While our approach mainly focuses on modeling what wiki pages are about, we also emphasize representation of the complete wiki server meta-data (wikis, users, pages, relationship between both ...) in RDF, as well as a way to interlink those two levels of representation. Moreover, it allows different wikis of the server to share information between each other while keeping a trace of any created information;
- *Immediate reuse of formalized data.* Finally, RDF data created among the wikis must be immediately reusable to enhance browsing and querying capabilities of the system, as in most of semantic wikis. More than adding value to the system, this is a way to show to users that their contribution has a real impact, which can motivate them to contribute regularly.

3.2 Global Architecture

To achieve the goals mentioned before, our architecture involves different components that are linked together, either logically or physically. The first part of the architecture consists in a set of ontologies that are used to model RDF data from the wikis, whether it is data about the pages or about their content. For the latter, ontologies have to be defined in RDFS or OWL and can be ontologies created specifically for the needs of a knowledge field, as well as existing ontologies. Regarding the RDF description of wiki pages and user actions, we mainly rely - and worked - on the SIOC ontology [16]. SIOC, which has recently been accepted as a W3C submission[2], provides a model to describe information about online communities and their content in a uniform manner and thus fits exactly in the scope of our needs. Moreover, it features a *type* module[3], so that we can represent wiki pages (`WikiArticle` class), but also wiki containers themselves (`Wiki`). We extended the SIOC ontology with a specific property, `embedsKnowledge` in order to formally represent in RDF the link between a wiki page and the data embedded in it (Fig. 2). This property provides a way to link any instance of `sioc:Item` - and its subclasses - to the URI of a named graph [17], i.e., in practice, the URL of a document that contains a set of RDF triples.

Then, the system features a web interface, based on a fork of the Drupal flexinode module[4], to edit wiki forms and browse and edit wiki pages. The editing interface combines plain-text and structured parts in order to easily manage the creation of RDF statements according to the ontologies it uses. When creating a wiki page, two documents are created, one containing the RDF description of the page (i.e. *"This is a wiki page created on 28th January 2008"*), the other one containg RDF data about its content (i.e. *"EDF is based in France"*). Both are linked thanks to the `embedsKnowledge` property we previously introduced.

The last component of the system is a triple-store, based on the 3store[5] API. By storing in real-time all RDF data as well as ontologies in a single place, it offers querying capabilities for the complete data and meta-data of the whole wiki-farm, but nevertheless keeps a trace of each statement thanks to its named graphs compliance. This store also manages basic inference capabilities and supports SPARQL and SPARUL (SPARQL update) in order to query and update data from the wiki pages.

Moreover, since all wikis of the wiki farm share the same knowledge base, by querying and updating a single RDF store, data can be re-used across wikis. Thus, an ontology instance created in a given wiki can be linked to an ontology instance from another one, even if there is not direct hyperlink between the pages that embeds this knowledge. It allows our system to create knowledge on a distributed way, even between various communities that do not share the same wiki but that produce information about the same ontology instances.

---

[2] `http://www.w3.org/Submission/2007/02/`

[3] `http://rdfs.org/sioc/types`

[4] `http://drupal.org/modules/flexinode`

[5] `http://threestore.sf.net`

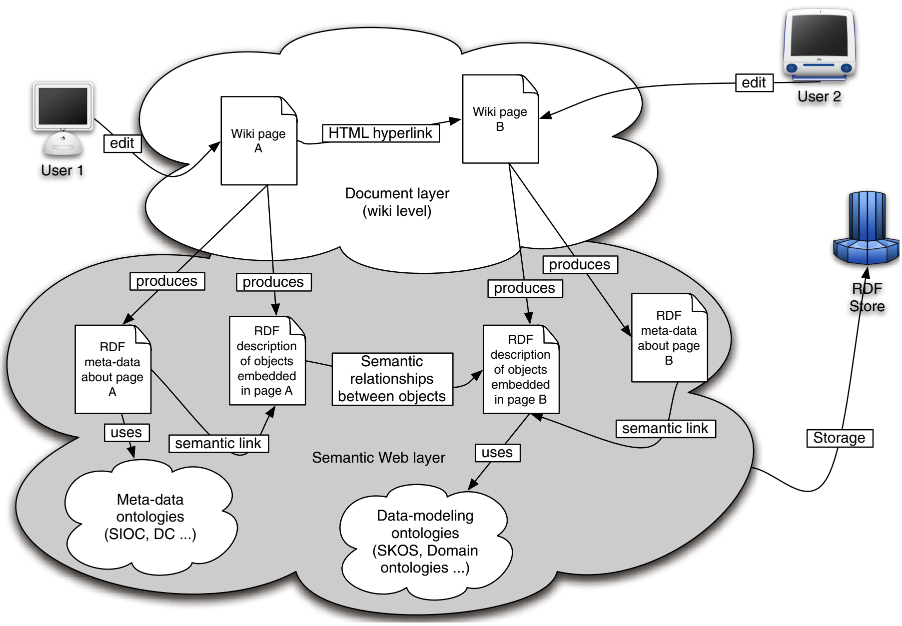


Fig. 2. Architecture of one wiki from the wiki-farm

4 Editing and Browsing Enhanced Wiki Pages

4.1 Manage Ontology Instances with Wiki Forms

As most semantic wikis, our system automatically creates ontology instances for each wiki page. While some wikis do not explicitly assign them a given type and other rely on the page category to identify it, our system requires administrators to define different types of pages, each one related to a class of a given ontology, e.g. a people page (related to `foaf:Person`[6]) or a software page (`doap:Project`[7]).

Regarding definition of properties and relationships of each instance, we use a mix of plain-text and forms in the wiki editing interface, thus separating plain-text content from content to be modeled in RDF, as the Semantic Forms extension[8] for Semantic Wiki or Freebase[9] do (Fig. 3). In our case, for each type of page, administrators define which form widgets they want to use and assign ontology relationship to each of them. When creating the page, translation from wiki content to RDF data is then automatically done thanks to those mappings. In a corporate context, we think that this combination of plain-text and forms to ease the modeling of RDF data has numerous advantages:

[6] http://foaf-project.org

[7] http://usefulinc.com/doap/

[8] http://www.mediawiki.org/wiki/Extension:Semantic_Forms

[9] http://www.freebase.com

- First, as fields are defined by the wiki administrators for each type of page and so for each class, users know what kind of knowledge is relevant for the wiki regarding a given page and can focus on essential aspects in this context;
- Moreover, as we kept a simple WYSIWYG field for each page, any other relevant information can be added, even if it has not been formalized as widgets. It can also help to participate in evolution of the model itself;
- Users can benefit from autocompletion features, suggesting possible related instances by querying the RDF store with on-the-fly SPARQL queries, thanks to AJAX technologies;
- At last, in our system, this approach allows to create complex relationships and ontology instances inside a single page. While most semantic wikis allow only to create relationships between existing objects, a widget can correspond to an dedicated class in our system, offering better ways to manage complex ontologies population.

While each page corresponds to a given ontology instance, instances are also created for each filled relationship field where a class has been assigned as a range. Then, if one later decides to create a wiki page for these instances, properties will be added to the existing ones. Moreover, when instances are not used anymore in any wiki, i.e. do not have any property, they are automatically removed from the RDF store to avoid orphan instances.

From these aspects, the wiki really acts as a collaborative ontology population tool, beneficing from Web 2.0 features to provide this task, since an individual

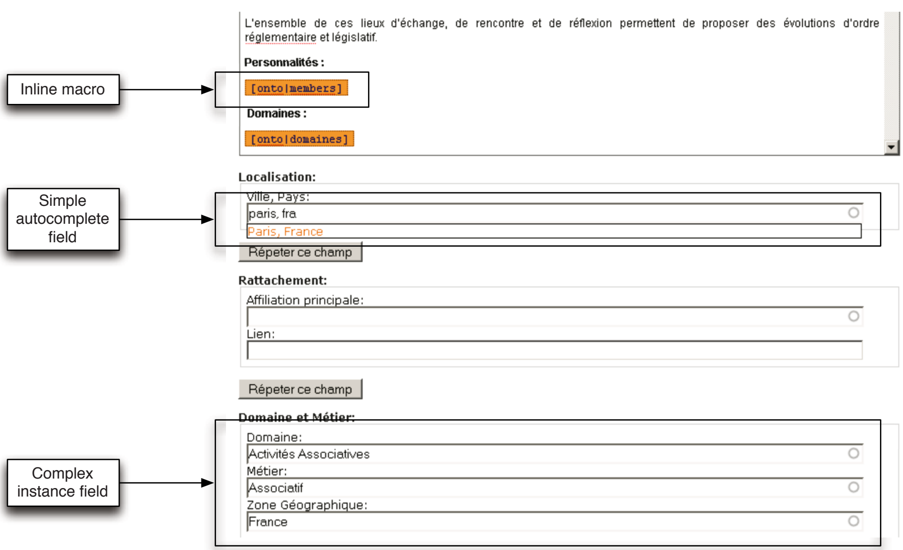


Fig. 3. Wiki editing interface

can be created by a user, modified by another, then linked to a third one by another one and even removed from the knowledge base by a fourth one.

4.2 Immediate Reuse of Knowledge When Browsing Wiki Pages

The main feature to enhance wiki browsing capabilities in our system is the use of *inline macros*, similar to inline queries of Semantic Mediawiki. Yet, those macros are defined by wiki administrators themselves, using SPARQL and PHP to render the results and are then called by users in wiki pages. Since all data are based on a set of predefined ontologies, queries can be written without having to deal with semantic heterogeneity problems, as people that would have use different property names for the same one, e.g. `isLocatedIn` versus `has_location`. The system then runs the query over the RDF store when the page loads, so that query results are always up-to-date. While queries can be complex, users simply type function names, with some arguments if needed, to use it in wiki pages. For example, `[onto|members]` will be translated in a query that will retrieve all people that are member of the organization described in a wiki page (Fig. 3, Fig. 4). Such queries take inference capabilities of the system into account, so that, for example, if they must list all organization instances described in the wiki, they will also lists companies or associations if they have been defined as subclasses of the first one in the ontology. Finally, the administrator can decide that the macro will render a link to add new page in the wiki to create an instance of a given type, thus facilitating the process of creating interlinked data.

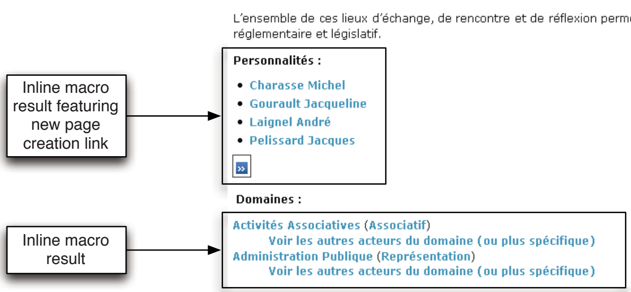


Fig. 4. Browsing an enhanced wiki page

Moreover, macros can take into account the way we combine modeling of data and meta-data in RDF export of wiki pages, so that a wiki can display a list of pages from another wiki for a given query. The snippet below shows a SPARQL query, that can be run as a macro from a given wiki and that will retrieve all pages (and their titles) that produced any RDF knowledge about EDF and that are contained in the wiki #6. This feature can also be used in the autocompletion of wiki forms, thus restricting the suggestions to data created from a single wiki, rather than suggesting data from the whole knowledge base.

```
select ?page ?title
where {
  graph ?data {
    :EDF ?predicate ?object
  } .
  ?page :embedsKnowledge ?data ;
    rdf:type sioct:WikiArticle ;
    dc:title ?title ;
    sioc:has_container <http://example.org/wiki/6> .
  <http://example.org/wiki/6> a sioct:Wiki .
}
```

Finally, those macros can display results according other rendering inter-faces, such as Google maps, in case the needed information is available in the RDF store. In order to achieve this goal, our wiki features an interface to query the GeoNames webservice[10] each time a new `geonames:Feature` instance is created in the system. Thus, while the result is similar to what can be done with the map view of OntoWiki, users do not have to manually enter the coordinates of each instance (e.g. a company) but simply fill a *"City, (State), Country"* field, that will be used to retrieve the appropriate RDF data - including coordinates - from geonames an add it in our knowledge base. Here, we clearly see the benefit of using the same model (i.e. the geonames ontology) as an existing RDF dataset to include data from external services at zero-cost.

5 First Experiments

We are currently experimenting this prototype at EDF R&D in a wiki that mainly deals with roles, locations and relationship of various agents involved in the energy field and its related domains. For this purpose, we build a set of lightweight but interlinked ontologies to model our data. One of this ontology extends the FOAF [18] vocabulary by defining new subclasses of its `foaf:Agent` class, as `Company` or `ResearchInstitute`, while another one consists in the definition of role of those agents, combining hierarchies of domains and role types (thanks to SKOS [19]) in a geographical context.

About 20 users are contributing to the wiki who are, for almost all of them, first-timers in the wiki world and so in the semantic wikis one. Thus, they felt quite comfortable with the wiki interface, especially thanks to the WYSIWYG part of the editing page and the non-technical side of its interface, which helps them to concentrate on the data without thinking on how to write or model and interlink the data. Yet, since the wiki itself deals only with ontology instances management and not with the model evolution, face-to-face meetings have to be organized between users and administrators to get their feedback and see what knowledge that they capture in free-text could be moved to the ontology with new forms, so that it can be re-used a better way.

[10] `http://sw.geonames.org`

So far, about 100 wiki pages and 200 ontology instances have been created. Actually, while some people focus on the RDF data modeling, other just concentrate on the WYSIWYG field, which is more intuitive to them. Unfortunately, it offer less possibilities to access their page, except direct hyperlink, while pages that produced RDF data can be linked from other pages thanks to the macro system we introduced before. The wiki features, nevertherless, lets some users update pages that have been started by others. The following graph (Fig. 5) offers a view about the growth of the knowledge base.

Finally, an important thing to notice is that we benefit from the ideas of architecture of participation and collective intelligence in this approach of collaborative ontology management. Indeed, users instinctively populate and maintain an ontology collaboratively, sometimes without knowing it, by simply using the wiki as they would usually do to reach their own needs by using such a tool, but offering a whole better experience of wiki browsing for any user of the system.

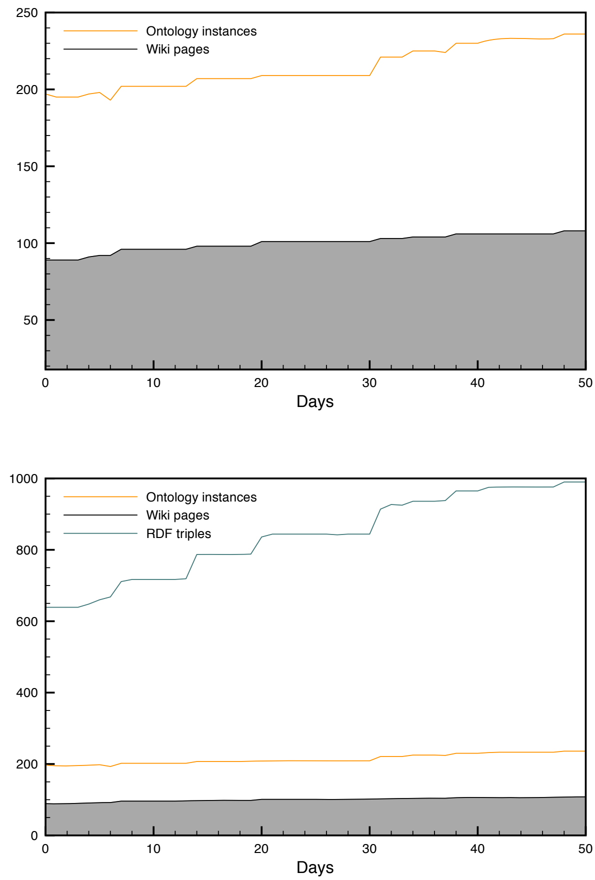


Fig. 5. Evolution of pages, instances and RDF triples on a 50-day period

6 Conclusion and Future Works

In this paper, we described a prototype of wiki that combine structure and Semantic Web modeling capabilities to produce ontology-based and machine-readable data in a collaborative way. We argued why we think that using structure in the wiki interface to produce this data can be a benefit in business information system regarding the adoption of the tool and the quality and reusability of the created knowledge. Finally, we showed how such knowledge can be used to enrich functionalities of the wiki. While this system combines some features that already exist in various prototypes, it focuses on usability for end-users, which is important in such a business information context, as well as, from the technical side, a common modeling of data and meta-data and ways to make various wikis share informations they help to build.

Regarding our future works, we will concentrate on adding new value-added functionalities to the wiki for end-users to ease the discovery of relevant information from the set of RDF data, as faceted browsing [20]. We will also focus on how to formalize wiki pages versioning in RDF, in order to see how statements about a given resource can evolve during its lifetime and track more precisely each change of information on a given ontology instance.

Acknowledgements

We would like to thanks the ID-Net team from EDF R&D for their input on the current experiments, as well as the reviewers of this paper for their valuable feedback.

References

1. McAfee, A.P.: Enterprise 2.0: The Dawn of Emergent Collaboration. MIT Sloan Management Review 47(3), 21–28 (2006)
2. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American 284(5), 34–43 (2001)
3. Ankolekar, A., Krötzsch, M., Tran, T., Vrandecic, D.: The Two Cultures: Mashing up Web 2.0 and the Semantic Web. Journal of Web Semantics 6(1) (2008)
4. Gruber, T.R.: Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In: Guarino, N., Poli, R. (eds.) Formal Ontology in Conceptual Analysis and Knowledge Representation, Kluwer Academic Publishers, Deventer, The Netherlands (1993)
5. Guarino, N.: Formal Ontology and Information Systems. In: Guarino, N. (ed.) Proceedings of the 1st International Conference on Formal Ontologies in Information Systems, FOIS 1998, Trento, Italy, pp. 3–15. IOS Press, Amsterdam (1998)
6. Bizer, C., Heath, T., Ayers, D., Raimond, Y.: Interlinking Open Data on the Web. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, Springer, Heidelberg (2007)
7. Berners-Lee, T.: Giant Global Graph (2008), http://dig.csail.mit.edu/breadcrumbs/node/215

8. Prud'hommeaux, E., Seaborne, A.: SPARQL Query Language for RDF. W3C Recommendation (January 15, 2008) (2008), http://www.w3.org/TR/rdf-sparql-query/
9. Tetlow, P., Pan, J.Z., Oberle, D., Wallace, E., Uschold, M., Kendall, E.: Ontology Driven Architectures and Potential Uses of the Semantic Web in Systems and Software Engineering. W3C Working Draft Working Group Note 2006-02-11 (2006), http://www.w3.org/2001/sw/BestPractices/SE/ODA/
10. Krötzsch, M., Vrandecic, D., Völkel, M.: Semantic MediaWiki. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 935–942. Springer, Heidelberg (2006)
11. Oren, E.: SemperWiki: A Semantic Personal Wiki. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, Springer, Heidelberg (2005)
12. Schaffert, S.: IkeWiki: A Semantic Wiki for Collaborative Knowledge Management. In: 1st International Workshop on Semantic Technologies in Collaborative Applications (STICA 2006) (2006)
13. Auer, S., Dietzold, S., Riechert, T.: OntoWiki - A Tool for Social, Semantic Collaboration. In: Proceedings of 5th International Semantic Web Conference 2006, pp. 736–749 (2006)
14. Buffa, M., Gandon, F.L., Sander, P., Faron, C., Ereteo, G.: SweetWiki: a semantic wiki. Journal of Web Semantics 6(1) (2008)
15. Auer, S., Bizer, C., Lehmann, J., Kobilarov, G., Cyganiak, R., Ives, Z.: DBpedia: A Nucleus for a Web of Open Data. In: Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC 2007), Busan, South Korea, pp. 715–728 (2007)
16. Breslin, J.G., Harth, A., Bojars, U., Decker, S.: Towards Semantically-Interlinked Online Communities. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532, pp. 500–514. Springer, Heidelberg (2005)
17. Carroll, J., Bizer, C., Hayes, P., Stickler, P.: Named Graphs, Provenance and Trust. In: Proceedings The Fourteenth International World Wide Web Conference (WWW 2005), Chiba, Japan (2005)
18. Brickley, D., Miller, L.: FOAF Vocabulary Specification 0.91. Namespace Document (November 2, 2007) - OpenID Edition (2007), http://xmlns.com/foaf/spec/
19. Miles, A., Brickley, D.: SKOS Core Guide. W3C Working Draft. Published online on November 2nd, 2005, (2005), http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102/
20. Oren, E., Delbru, R., Decker, S.: Extending faceted navigation for RDF data. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, Springer, Heidelberg (2006)

Agreeing While Disagreeing, a Best Practice for Business Ontology Development

Emanuele Della Valle, Irene Celino, and Dario Cerizza

CEFRIEL – Politecnico of Milano, Via Fucini 2, 20133 Milano, Italy
{emanuele.dellavalle,irene.celino,dario.cerizza}@cefriel.it

Abstract. The agreement is a crucial part of our living together. Important opportunities for sharing resources, integrating systems and collaborating depend on our ability to agree. While we are interested in methods and technologies that support shared agreement, we somehow tend to forget the disagreement; indeed, it is also a part of reality.

In the Semantic Web field, most of research activities investigate ontology agreement and its formalization, rather than the disagreement and the best practices about its extended use.

We claim that the industrial uptake of the Semantic Web is severely limited by the fact that, in the real world, shared agreements are difficult to reach and maintain and that "agreeing while disagreeing" is a common practice. In this paper we illustrate a best practice, which we learned from practical experience, that makes use of an (almost) unexplored potentiality of the Semantic Web to express disagreement, and we describe its use in the e-employment sector.

1 Introduction

Whenever we agree with someone, we feel free to express our consent (*"I think you're right"* or *"I agree with you"*), because the agreement does not imply any conflict. On the other hand, whenever we have to dissent from someone, we feel embarrassed to express our disagreement (*"I'm afraid I can't agree with you"*) and we often try to find the most polite expression in order not to offend our interlocutor (*"I agree with you up to a point, but..."*, *"That's quite true, but..."* or *"I agree with you in principle, but..."*, etc.). As a consequence, every spoken language provides for a lot of expressions to formulate (or to hide) dissent and a fewer expressions, that we daily use without any worry, to state accordance and harmony.

It seems that we fear both the disagreement and the conflicts that may rise by showing it to the others. After all, we can happily work together by committing to a limited agreement, while hiding all the necessary (and unnecessary!) discordances. Such behavior is theoretically founded in the *common knowledge* theory [1], which demonstrates that, when showing the whole agreement, parties cannot "agree to disagree". Therefore, the only way to disagree is to show only part of the agreement. It may appear counter-intuitive, but a good and common practice is to "agree while disagreeing".

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 70–82, 2008.

What happens when you have to develop a Semantic Web application that deals with such a common behavior? Since an accepted definition of ontology is a "formal, explicit specification of a *shared* conceptualization" [2], we can clearly assume that an ontology is a kind of formal and explicit agreement among the parties that commit to it.

Is there room for disagreement? Being able to share the disagreement may not be a desirable feature, because we tend to agree on many general principles, to commit to a clear subset of the possible implications of this principles and, in the meantime, to keep our disagreement. After all, however, disagreement exists.

Our thesis is that the Semantic Web is elaborating all necessary tools to formalize the agreement and the disagreement, and that their appropriate use can result in a method that helps in developing real-world Semantic Web applications.

The rest of the paper is structured as follows: section 2 states the problem of "agreeing while disagreeing" and presents the conceptual solution we elaborated in several research projects; the notion of ontological commitment is introduced in section 3. Section 4 draws a boundary line between agreement and commitment. Section 5 explains in details our best practice for expressing disagreement and making it partially public; section 6 outlines the method we put in practice in elaborating the SEEMP solution and its advantages; finally section 7 presents conclusions and future work.

2 The Problem and a Conceptual Solution

We have been technically leading several European projects (i.e., COCOON[1] and SEEMP[2]) and Use Case work packages of European and Italian projects (i.e., in TripCom[3] and NeP4B[4]) that aim at fostering the adoption of Semantic Web in health care, government and business. In all those projects, we face the problem of integrating, using a Semantic Web approach, data and applications that belong to parties that are definitely interested in collaborating, but are also competitors. Therefore, they have to agree about the sharing of all the information that will foster collaboration, but they also should not be forced to disclose information that will reduce their competitive advantages.

In those cases, parties agree on general principles, but then they only commit to a subset of the possible implications of these principles and keep for themselves all their disagreements.

For instance, in the domain of e-employment (scenario of the SEEMP project), different Employment Services, both public and private, collect CVs and job offers in order to match demand with supply. Each Employment Service (ES) covers either a region or an occupational sector. As a result, the employment market is severely fragmented and many ESs perceive the need of sharing information in order to provide a better service to their customers. However, they

[1] http://www.cocoon-health.com/
[2] http://www.seemp.org/
[3] http://www.tripcom.org/
[4] http://www.dbgroup.unimo.it/nep4b/

would never exchange CVs or job offers, since they contain sensible information (like contact details); instead, the ESs use the "anonymized" versions of CVs and job offers, which we name respectively candidacies and vacancies. Therefore, if an ES exchanges a candidacy/vacancy with another one, it potentially enlarges the possibilities of finding a match, without giving to the other one the chance to by-pass it and directly contact the employee/employer.

In a similar manner, in the logistic domain (scenario of the NeP4B project), different logistic operators may join their efforts to acquire a larger market share by presenting themselves under a common brand. They make available information on their number and type of vehicles, their position on the territory, etc., but, being competitors, they only partially disclose their respective price lists. A joint logistic platform should have enough information to calculate the price of a shipment request, but not enough to let one of the logistic operator to calculate a competitive counter-offer.

Therefore, we draw the conclusion that the common notion of *shared agreement* is not enough to manage the complexity of the industrial scenarios we face. We believe that two important notions have to be made explicit: *commitment* and *disagreement*. It is worth noting that usually when parties provide ontology commitment the intended meaning is that all parties commit to the "entire" ontology. On the contrary, we propose to give a "subjective" meaning to commitment and disagreement which does not presume a common knowledge among all the parties: two parties may commit to (or disagree with) different part of the agreement.

In order to move from the problem statement to the solution conception, we need to find appropriate methods and technologies. We need conceptual elements to capture the notions of agreement, commitment and disagreement, to make them operational and to express the respective relations among them.

Ontologies have been used and are good for formalizing and sharing the agreement. The notion of commitment is usually associated to the notion of ontology, and this is certainly true in the context of agent communication (see also section 3). In agent-based systems, all agents usually share a single ontology. The Semantic Web vision, however, foresees an ecosystem of ontologies, because of the very nature of the Web which is "fractal" [3]. Ontologies can be co-invented, they can partially overlap and, in developing a new ontology, the importing of existing ones is encouraged [4]. We believe that the "practical" meaning of ontological commitment in the Semantic Web is slightly different from the original one (see also section 3). In formal terms, committing to an ontology that imports several other ones is the same as committing to one big ontology obtained by the union of all of them; however, in practical terms, committing to the ontology that includes the import annotations is partially an "unconscious" commitment, in that it means trusting the knowledge engineer who decided which ontologies to import.

Therefore, our best practice is to distinguish between the *reference ontology*, which captures the shared agreement, and the *local ontologies*, which captures the commitment and the disagreement of the various parties. We propose to build the reference ontology including all the details that are needed to allow for a meaningful communication between each pair of parties, thus including details that most of the parties would consider either inessential or not sharable. Then, each party can develop its local ontology, partially by importing parts of the reference ontology, and partially by developing its own point of view. Every time a local ontology imports a part of the reference ontology[5], the party is considered to commit to the imported parts of the reference ontology. Moreover, every time a part of the local ontology is aligned to a part of the reference ontology (e.g. by the means of ontology-to-ontology mediators [5]), the party is also said to commit to that part of the reference ontology. A particular attention should be given in capturing also the source of disagreement within the local ontology. Finally, each party should make available to all other actors the part of the local ontology that explains its different point of view without causing conflicts.

Sections 3-5 are devoted to explain the details of our conceptual solution.

3 Ontological Commitment

In the philosophy of language and metaphysics, the *ontological commitment* of a proposition is the definition of some of the demands that the sentence's truth imposes on the world, e.g., the existence of one thing presupposed or implied by asserting the existence of another one.

In the context of artificial intelligence, the ontological commitment is a direct consequence of the ontology definition: an ontology is a specification of the concepts and relationships that can exist for a community of agents and the ontological commitment is the agreement to use the shared vocabulary defined by the ontology in a coherent and consistent manner [6,7,8].

From the previous definition, it follows that an ontological commitment is a guarantee of *consistency*, but not *completeness* with respect to queries and assertions [9]: the agents sharing a vocabulary do not need to share a knowledge base; each of them knows things the others do not, and an agent that commits to an ontology is not required to be able to answer all queries that can be formulated in the shared vocabulary.

Also in the Semantic Web community, we often refer to the ontological commitment in the sense explained above. We are used to build ontologies to express the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them; the resulting ontology represents an abstract, simplified view of the world that we wish to formalize for some purpose. As a consequence, every knowledge-based system built on top of an ontology is committed to its conceptualization. This means that, when two different systems commit to the same ontology, they can assert the existence of different things and individuals (*in-completeness* of the commitment), but they share the same vision of the conceptual model expressed by the ontology (*consistency* of the commitment).

[5] Please note that the behavior of this import feature differs from the behavior of `owl:import` annotation, because it allows for a *partial* importing. However it can be mapped to `owl:import` is the reference ontology is correctly modularized.

Several different methodologies developed to build ontologies [10] are based on this notion of ontological commitment, in that they assume the ontology engineering process to be aimed at formalizing the obligation to a specific conceptualization of the world (or of a part of it). As we mentioned in section 2, the only difference on the notion of ontological commitment between the Artificial Intelligence and the Semantic Web world seems to be that, in the latter case, the engagement to the ontology is explicit and conscious with regards to the concepts, relations, axioms and individuals defined within the ontology itself, while, in the Semantic Web, it is somehow unconscious with respect to the ontologies imported via the `owl:import` mechanism.

4 Agreement vs. Commitment

In proposing our best practice, we would like to stress the difference between the achievement of an agreement and the commitment to the obligations it entails.

To better explain this distinction, we take as an example the situation in which there are three different actors wishing to pave the way for a fruitful collaboration between them. To start their interaction, they need to formalize the common domain of discourse. If two of them need the specification of some concepts on which the third actor disagrees, they have two possibilities:

1. either ignoring the disputed concepts; this results in a poor and weak agreement that prevents the two concordant actors from fully collaborate;
2. or including the controversial concepts; this results in a stronger agreement that however can heavily hinder the relationship with the third actor.

We believe that the *shared agreement* must be large enough to allow for a *meaningful communication* between all the three parties to be integrated. It can include details that are needed by two parties and inessential for the remaining one. However, with this extended meaning of shared agreement, the third actor will refuse to commit directly to the *entire* agreement, because it includes details that are useless from its point of view.

Our best practice proposes to solve the problem by allowing for a *partial commitment* to the shared agreement, which means that each party engages itself in a durable obligation of complying with just a part of the agreement. Parties that need details will commit to the part of the agreement that contains them, whereas those that do not need them will not.

Therefore, to preserve the common negotiation and to enable, as much as possible, a win-win situation between all the parties of a business deal, we suggest to *keep the separation between the agreement and the commitment to it*. In other words:

– by *agreement*, we mean the description of a common field of action, potentially going in all the details which enable a meaningful communication between at least two parties of the arrangement; while,
– by *commitment*, we mean the specification of (the part of) the agreement on which an actor engages itself by complying with it; the commitment

therefore implies an active and durable obligation to a part of the agreement, while preserving the party's reservation and disagreement on the rest of the agreement itself.

We distinguish from the definition of ontological commitment given in [9] in that we extend the notion of in-completeness of the commitment from the factual knowledge to the conceptual knowledge: each party (or agent) is free to discard parts of the ontological agreement and to undertake to the sharable portion only.

As mentioned in the section 2, the realization of our definition of agreement relies in the explicit conceptual formalization of the *reference ontology*. This means that, whenever three or more parties settle an arrangement and reach a shared agreement, they formalize it in the reference ontology. The ways to (partially) commit to the reference ontology are detailed in section 5.

5 Preserving the Disagreement

The commitment to the agreement is the way to express the accordance and the dissent with regards to the mutual consent. Therefore, we need a way to explicitly express the disagreement of a party on the object of the common action field.

We believe that, in order to preserve the disagreement, it is useful to consider two orthogonal dimensions of the commitment: positive vs. negative commitment and public vs. private commitment. A final summary of these notions is displayed in table 1 on page 75.

Table 1. The "traditional" notion of commitment can be refined in four different notions, to deploy real Semantic Web application that enable meaningful communication among parties while leaving crucial leeways of disagreement

Our Best Practice in Preserving the Disagreement

- *Public Positive Commitment* is the space of "traditional" commitment (as proposed by T. Gruber [9]), which allows for a meaningful communication with other parties.
- *Public Negative Commitment* is the space for disclosing knowledge about well known disagreement that will not cause conflicts among the partners, but that, on the contrary, allows for checking in advance if asking queries to a party is worth a try.
- *Private Positive Commitment* avoids a party to let every other parties know that it committed to answer certain queries; however, it leaves the possibility to ask queries to a subset of partners that are informed about the positive commitment (by using a different channel).
- *Private Negative Commitment* enables the possibility to write filters that prevent "illegal" information to be exchanged, without disclosing knowledge that may cause a party to come into conflict with other ones.

5.1 Positive vs. Negative Commitment

We say that a party expresses a *positive commitment* whenever it imports a part of the reference ontology or aligns a part of its local ontology with the reference one. Positive commitment provide a means for meaningful communication on the basis of the parts of the reference ontology that received commitment by other parties. Such parties, thus, can ask queries to the positive-committed party and they may expect answers whenever the knowledge base of the party contains useful facts.

On the contrary, a party expresses a *negative commitment* in two ways. The easiest one is by avoiding to express a positive commitment on a part of the reference ontology. However, this way may not be enough: the second way consists in formalizing parties' specific points of view (i.e., the source of disagreement).

The party that *refrains from expressing a positive commitment* on a part of the reference ontology prevents other parties to establish meaningful communication that involves that part of the reference ontology. When other parties ask queries about such part of the reference ontology, the party will not answer, because it does not understand the query.

However, the same party can do more than refusing to commit, it can *formalize its point of view* as a part of its local ontology. As a result, one party can refine its positive commitment, which is supposed to include leeways for disagreement, by expressing its specific point of view.

For instance, in the case of employment, each European country has its own regulation in terms of job legislation and skills/certifications required to apply for a job. This is especially relevant for physician, lawyers, teachers, etc. Those regulations are mandatory for each country, but, being "local", they cannot fall within the shared agreement (i.e., the reference ontology).

As a concrete example, let's consider two Employment Services (ES), one Swiss and one Italian. Both ESs express a positive commitment on the concepts related to University occupations. However, the legislation about the prerequisites to apply for a University professor position is different between Switzerland and Italy: the two countries disagree on the necessity of holding a Ph.D. title. Therefore, the Swiss ES also makes explicit in its local ontology that each candidate for a professor position should hold a Ph.D. title (whereas in Italy this is not mandatory).

The negative commitment, under the form of an explicit formalization of the local peculiarities, can be used by the discordant party, in order to actively filter the queries it receives.

In the example above, the Swiss ES can write a rule that filters candidates that do not hold a Ph.D title. So, if the Italian ES asks for available University professor positions for an Italian candidate that did not get a Ph.D. title, the Swiss ES will not provide any matching vacancy, even if one is available.

5.2 Public vs. Private Commitment

We say that a party makes its commitment *public* if it discloses to the other parties involved in the agreement its positive/negative commitment. It is worth noting that making the entire commitment public is not mandatory.

We believe that the positive commitment should be always made public, because it lets other parties know in advance if it is worth asking a query (this can be done by checking if the query involves the part of the reference ontology which the receiver commits to). However, a party may decide to only *partially* revealing its positive commitment to all the participants, because it prefers answering only to a subset of them; in this situation, the party discloses its ability to answer certain queries only to the "friendly" parties and by using a different communication channel.

In a similar manner, a party may decide to disclose only part of its negative commitment. Making public a negative commitment is particularly meaningful if it communicates a well known disagreement (e.g., different legislations among EU countries). Such negative commitment can be evaluated by other parties before establishing a collaboration, in order to prevent formally invalid business processes to be activated. However, keeping private large part of the party's specific point of view is quite usual, because by revealing it the party can come into conflict with others.

Such *private* part of the negative commitment can also be used to express filters that prevent "illegal" information from being exchanged. Let's consider for instance an Italian logistic operator (LO) which is required to transfer a load of soft drugs. Being such request illegal in Italy (but legal in some other countries where the request could have been started), the LO should not only refuse to process the shipment order, but also prevent the request from entering into the elaboration system.

5.3 Language and Tool Support

In putting into practice what we discuss above (see also section 6), we were severely limited by the lack of support in standard languages (RDF, RDFS and OWL) and tools.

Import annotations in OWL are not flexible enough; a partial workaround would be a careful modularization of the reference ontology. However, a full implementation of our approach would require an importing clause that enables to select a part of the ontology.

Moreover, in order to express disagreement, we need both a rich ontological language for expressing inequality and a rule language for describing ontology alignments and formalizing filters on data exchange. For what regards the former, we notice that the current support for inequality expression in OWL is limited to `owl:differentFrom` and `owl:disjointWith`; the latter requisite is partially satisfied by different proposals, but no standard Semantic Web rule language exists yet.

For those reasons, we look forward to following the standardization process of OWL 1.1 [11], which includes several new features that we consider interesting for our purposes (e.g., the application of disjointness not only to classes but also to properties and the possibility to predicate negative assertions). In a similar manner, we expect RIF Core [12] to be of help; however, the tool support will probably be limited until the standardization process is in progress.

Using best-of-breed solutions, in the SEEMP project we adopted the WSML language [13] and WebODE [14] and WSMT [15] as tools. WSML offers a layering of semantic languages including a rule language (WSML-rule), which can be used to describe ontology alignment and to state complex axioms (such as those required to describe filters), on top of an ontological language (WSML-light), which can be used to model the declarative part of the ontologies. WebODE and WSMT were chosen because they provide complete support for WSML, respectively, to model ontologies and describe alignments and filters.

6 The SEEMP Solution

In the following, we present our approach in developing the core of a Semantic Web application that allows parties to partially disagree while doing business together. As explained in details below, our approach mainly comes from the experience we gained during the SEEMP project. The lesson learned and the best practice we would like to share is outlined in table 2.

Table 2. Our recipe to express agreement and disagreement

The Way We Applied Our Best Practice in the SEEMP Project

To formalize agreement and disagreement, we use:

- *Ontologies* as way to express:
  – the general shared agreement among all the parties, which we name *reference ontology*; and
  – the specific points of view of the parties that are only partially sharable with others, which we name *local ontologies*.
- *Ontology-to-Ontology Mediators* as a way to express:
  – the *commitment* toward the reference ontology by (partially) aligning the local and the reference ontologies; and
  – the formal expression of the *disagreement* of the party, by defining filters that exclude the possibility for facts expressed in the local ontology to be translated in the reference form or vice-versa.

Moreover, with regards to the respect of each party's competitive advantage or private data, we suggest:

- to *make publicly available* to all parties the reference ontology and the parts of the local ontologies and the respective mediators which express the local point of view and which is not in contrast with the reference ontology;
- to *keep private* to each party the disagreement expressed in the rest of the local ontologies and the respective mediators.

We successfully experiment the previously described approach in the SEEMP project, in which concretely we have:

– the *SEEMP Reference Ontology*, as the common agreement between the various ESs that take part in the employment marketplace;
– the various *SEEMP Local Ontologies*, which are the specifications of the local agreements, i.e. the formalizations of the peculiarities of the local environments;
– the *SEEMP Connectors*, as the mediators between the real-world information systems and the shared agreement; they represent both the "positive" commitment onto the Reference Ontology and the "negative" commitment, i.e. the formulation of the specific content or needs of each local ES.

The *Reference Ontology* is a core component of the SEEMP system: it is the common "language" to describe the details of the employment sector. It is rich enough to support the semantic needs of all the ESs, currently involved or subsequently integrated in the SEEMP marketplace. The Reference Ontology, as well as the Local Ontologies, have been developed by following the Methontology approach [14]; this serves to have scalable, adaptable and maintainable ontologies.

For what regards the *Local Ontologies*, in the SEEMP architecture we had two possible options to build them:

1. taking as a seed the Reference Ontology: in this case, the concepts in the Local Ontologies are extension in depth of the concepts already present in the Reference Ontology; the consequence is that the data exchange between different ESs is easy, while the complexity of the Connectors between the Local Ontologies and the local schemata is higher;
2. operating a reverse engineering process from ES schema sources: it is the easiest way for ontologizing ESs, since each concept in a Local Ontology is the semantic expression of a relevant concept in the respective ES; the consequence is that the Connectors between Local Ontologies and schema sources are not complex, while the mappings between the Reference and Local Ontologies can be difficult and cause delay in the data exchange.

The SEEMP way adopts the first option at the beginning, when few ESs are present in the marketplace, while when more ESs join the marketplace, the solution will move toward the second option. The balance between the two options is related to the need for a "minimal shared commitment".

Since each ES talks in its own "language" (the Local Ontology, which represents its view on the employment domain), its respective connector is responsible for *resolving* these *heterogeneity issues*, by translating the local content in terms of the Reference Ontology. In this way, all the ESs in the marketplace speak the same language, and heterogeneity issues are solved, because, rather than managing mappings between every possible ontology pair (which becomes unmanageable as the number of ESs grows), each ES need only maintain mappings to and from the Reference Ontology.

The SEEMP approach offers several advantages, which can be analyzed from the point of view of both CEOs (the decision makers) and CTOs (the IT experts).

The SEEMP solution offers CEOs a way to enforce the subsidiarity principle [16], by valuing each ES contribution in the marketplace. This increases the number of interconnections, which in turn results in a larger number of faithful users and business transactions.

From the point of view of CTOs, SEEMP enables an easier maintenance of the integration with other ESs and lower integration costs. Semantics makes the mappings between different terminologies easier, because tools like WSMT [17] can analyze Reference and Local Ontologies (e.g., by comparing sub-structures and by searching for synonymies) and can guide the IT Administrator in drawing the mappings. Thank to this support, the mapping definition process requires less time or, alternatively, it provides more precise mappings in the same amount of time. This results in a lower integration cost.

There is also an added value from the point of view of the final users of the employment marketplace, i.e. job seekers and employers. Many job offers, that today could only be found by inserting multiple times the CV in each ES and by merging manually the results of different ESs, becomes seamlessly available through the interface of each ES.

In conclusion, to achieve all these results, what has to be built is a comprehensive Reference Ontology that encompasses several employment domains. Development and maintenance of this reference part of SEEMP is more then an ICT problem: it is a matter of reaching agreement at organizational level. As already discussed, the goal of SEEMP is reaching a "minimal shared commitment" in which ESs *agree* on high-level aspects, allowing for collaboration among them, *while disagreeing* on minor details that differentiate each ES from the others.

7 Conclusions and Future Work

During these years, the interest in interoperability issues shifted from a merely technical problem (how to develop Internet applications) to a business problem (how to create value by offering cross-enterprise services to the market).

As a consequence, new constraints arose. When a service is offered by a constellation of actors, a central authority is often difficult to set up and all parties aim at minimizing its role. The central authority should perform only those tasks which cannot be performed effectively by the parties. This principle is known as *subsidiarity* [16]: it states that matters ought to be handled by the smallest (or, the lowest) competent authority. Each party acts independently and a central system (if any) performs only operations that exceed the capacity of individual parties. Subsidiarity assumes that each party is interested in fostering collaboration to create a common value, but it saves room for competing with other parties for market share. Therefore, the principle of subsidiarity emphasizes the importance of an interoperability infrastructures, which supports and empowers each party's actions and which strengthens links among the parties in the market.

Contemporary solutions to interoperability issues need to protect different positions and ideas. They should take into consideration that looking for a comprehensive agreement sometimes can be useless or counter-productive, because the practice to exclude possible causes of conflict from the agreement can produce agreements that are so limited to be almost empty, thus useless. Last but not least, the explicit formalization of disagreement, if well-known and conflict-safe, is needed in most business domains, in order to prevent invalid business transactions from taking place.

In this paper, we proposed a best practice that could help in guiding the deployment of interoperability solutions based on Semantic Web which, in our opinion, has a better chance for industrial uptake. It proposes the formalization of the notion of agreement, commitment and disagreement. It recommends the development of a reference ontology, which is the common source of agreement, and several local ontologies, which express commitment and make the disagreement explicit. The adoption of our best practice in the SEEMP project to the employment market gave us several positive feedbacks and we are currently working in applying it in the NeP4B project to the logistic market.

We believe that we will be able to develop in the near future a full-fledged methodology and a comprehensive tool set, that support several of the tasks that now we can perform only manually.

Part of the method that we are formalizing consists in the iterative construction of both reference and local ontologies, based on the analysis of the public positive feedbacks of each partner. Starting from a basic reference ontology, parties provide rough public positive feedbacks. By formally evaluating the part of the agreement that received the commitment of every partner, it is possible to understand the level of *minimal* meaningful communication that can be established among all the parties. At the same time, each partner can understand which part of the agreement it has to commit to, if it wants to achieve a higher level of communication with one or more parties. After this analysis, the un-committed part of the reference ontology can be eliminated and parties are asked to provide again public positive commitment. After each iteration, the sources of conflict in the reference ontology are reduced and a broader positive commitment is reached. In the end, each party is invited to make public the negative commitment that cause no conflict.

We need to make the method formal, to develop supports tools, especially visualization ones, and to asses its usefulness in several projects. The refinement of the SEEMP solution and the development of NeP4B prototypes will be of great importance in the next years.

Acknowledgments

This research has been partially supported by the SEEMP EU-funded project (IST-4-027347-STREP) and by the NeP4B Italian-funded FIRB project (MIUR-2005-RBNE05XYPW). We thank all the colleagues, and in particular Asunción Gómez-Pérez and Gang Zhao, for the fruitful discussion.

References

1. Aumann, R.J.: Agreeing to Disagree. The Annals of Statistics 4(6), 1236–1239 (1976)
2. Studer, R., Benjamins, V.R., Fensel, D.: Knowledge engineering: Principles and methods. Data & Knowledge Engineering 25(1-2), 161–197 (1998)
3. Berners-Lee, T.: The Fractal Nature of the Web, working draft (1998–2007), http://www.w3.org/DesignIssues/Fractal.html
4. Shadbolt, N., Berners-Lee, T., Hall, W.: The Semantic Web Revisited. IEEE Intelligent Systems 21(3), 96–101 (2006)
5. de Bruijn, J., Ehrig, M., Feier, C., Martíns-Recuerda, F., Scharffe, F., Weiten, M.: Ontology Mediation, Merging, and Aligning. In: Semantic Web Technologies: Trends and Research in Ontology-based Systems, pp. 95–113. John Wiley & Sons, Ltd., Chichester (2006)
6. Newell, A.: The knowledge level. Artificial Intelligence 18(1), 87–127 (1982)
7. Guarino, N.: Formal ontology and information systems. In: Proceedings of the 1st International Conference on Formal Ontologies in Information Systems. FOIS 1998, pp. 3–15. IOS Press, Amsterdam (1998)
8. Mallya, A.U., Huhns, M.N.: Commitments among agents. Internet Computing 7(4), 90–93 (2003)
9. Gruber, T.: A translation approach to portable ontology specifications. Knowledge Acquisition 5, 199–220 (1993)
10. Nagypál, G.: Ontology development. In: Studer, R., Grimm, S., Abecker, A. (eds.) Semantic Web Services, Springer, Heidelberg (2007)
11. Patel-Schneider, P.F., Horrocks, I., Motik, B.: OWL 1.1 Web Ontology Language: Structural Specification and Functional-Style Syntax. Technical report, W3C (2006)
12. Boley, H., Kifer, M.: RIF Core Design. Technical report, W3C (2006)
13. de Bruijn, J., Lausen, H., Polleres, A., Fensel, D.: The web service modeling language: An overview. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, Springer, Heidelberg (2006)
14. Gómez-Pérez, A., Fernández-López, M., Corcho, O.: Ontological Engineering. Springer, Heidelberg (2003)
15. Kerrigan, M.: D9.1v0.1 Web Service Modeling Toolkit. Technical report, University of Innsbruck (2005)
16. 2004/387/EC: Decision of the European Parliament and of the Council on Interoperable Delivery of pan-European Services to Public Administrations, 2004 (2004)
17. Mocan, A., Cimpian, E.: Mappings creation using a view based approach. In: MEDIATE 2005. CEUR Workshop Proceedings, CEUR-WS.org, vol. 168, pp. 97–112 (2005)

Modeling and Using Polish Legal Knowledge - Commercial Companies Code Ontology

Piotr Stolarski and Tadeusz Tomaszewski

Poznan University of Economics, Department of Information Systems, al. Niepodleglosci 10, 60-967, Poznań, Poland
{P.Stolarski,T.Tomaszewski}@kie.ae.poznan.pl

Abstract. This paper is motivated by three goals: more advanced knowledge modeling techniques on the basis of the statute-specific legal ontology in the domain of commercial law are presented; secondly, a number of real-life problems solutions using a kind of reasoning system show how the reasoning may be held over the ontology; finally, we point out some difficulties which may be encountered at the building process of similar knowledge models. The ontology describes fundamental concepts taken from the Polish Commercial Companies Code which has been amended in order to comply Polish law with EU standards.

Keywords: Legal reasoning, law and computing, legal ontologies, Commercial Companies Code, AI and law.

1 Introduction

There has been a number of works and ongoing efforts in the field of legal knowledge modeling. A thorough reference of works related to legal ontologies may be found in [2]. Previously we demonstrated the mechanism of simple reasoning over the basic parts of the legal knowledge constructs [13].

In this paper we present the legal ontology of Polish Commercial Companies Code[1] (PCCC). In addition to this, some details about its architecture are discussed and two robust case examples of doing inference tasks as the practical use of our construction are shown. Our ontology represents PCCC concepts and relationships among them. Using the OWL DL[2] description language and the Racer Pro[3] reasoning system as the inference engine we are going to attain a twofold goal, namely: we present the possibilities of constructing statute-specific legal ontology in the domain of commercial law; secondly, we depict how some basic real-life problems may be solved using a powerful reasoning system combined with knowledge codified in our ontology; additionally, we outline some difficulties which may be encountered at the building process of similar constructs. In the course of our research the Protégé[4] [1, 4] ontology editor was used as the development tool.

---

[1] Polish name of the statute is "Kodeks Spółek Handlowych".
[2] http://www.w3.org/TR/owl-features/
[3] http://www.racer-systems.com/
[4] Ontology Editor and Knowledge Acquisition System, http://protege.stanford.edu/

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 83–94, 2008.

The paper is divided into four main sections which are as follow: 2. Legal ontologies – is the general introduction to the legal knowledge models, together with a comprehensive ontology definition; 3. The corpus – is a short description of the knowledge sources we employed during conceptual analysis of the ontology; in (4.) The reasoning section demonstrates detailed examples of reasoning about the concepts of the ontology.

2 Legal Ontologies

Legal ontologies are a major part of legal knowledge modeling efforts [2, 14]. The more recent ones use ontologies as a way of expressing the common understanding of concepts and relations among them in a formal and structured manner. Such knowledge repositories may be a subject to reuse (by inclusion or expansion) in a variety of information or intelligent systems. Moreover the employment of ontologies for the purpose of legal knowledge modeling enables large-scale machine processing as well as the access in a human-readable form via some user interfaces. The other aspect of formal knowledge specifications is the normalization of terminology. A normalized terminology poses much less threat of ambiguity in any communication or data exchange process. To mention only some more extraordinary possibilities that the technologies promise, we may enlist:

- Creating a common gateway facilitating information exchange between domains (i.e. taxes, property administering).
- Delivering a tool for checking and correcting of the formal side of day–to–day decisions.
- Facilitating tasks of disseminating of information about corporate entities among different administration levels as well as on cross-country level.
- Facilitating tasks of management of knowledge on companies, SME and other legal subjects.

The more comprehensive enumeration of legal knowledge models applications can be reached in [13].

Our model is a kind of a domain specific ontology with specialized legal vocabulary as we limit ourselves only to the definition of lexems corresponding to one statute, thus our legal ontology is statute-specific provided that we outsource (import) some extensions from other ontologies.

2.1 Ontology Definition

As mentioned earlier more recent efforts in the field of legal knowledge modeling treat ontologies as a method of unification of understanding of concepts and relations as a formal structure. The base definitions for our notion of what ontology is, is that given in [3]. The last one itself is derived from the OKBC knowledge model.

The Description Logic (DL) is a set of languages devised in order to represent knowledge in a handy and compact collection of simple facts. Since the DL formalisms are in fact a part (a subset) of a first-order predicate logic they are characterized with well structured syntax and formally explainable semantics. Such features are very desirable especially in the area of knowledge encoding. Therefore a number of ontology applications use DL to expand their simple descriptive capabilities. This gives us much more powerful possibilities of applying different kinds of reasoners, processing the contained knowledge.

Although DL languages may have different expressivities a minimal set of features can be defined. A typical syntax together with its semantical interpretation of the DL system is presented in [5].

OWL is a language devised for describing complex and modularized ontologies for the purpose of use in the Semantic Web. The main features of the language are summarized in [6].

3 The PCCC Ontology Corpus

3.1 Polish Commercial Companies Code

The main source of the corpus for the PCCC ontology is the Polish Commercial Companies Code Act (published in Journal of Laws No 94, item 1037). The Act was adopted by the Lower House of the Polish Parliament on 15th September 2001, and – with changes introduced later on - is valid legal source from 1st January 2001.

There were two main reasons of incorporating the new Act replacing the old Commercial Code of 1934:

- Improvement of the law quality by providing a new text coherent with all other legal sources.
- Assimilating the EU law into the Polish legal system.

Of course, during the work on the ontology we processed many legal texts and the most important legal acts. We limit ourselves to presenting a few excerpts [7] important for the rest of the discourse.

Art. 1. [Scope of regulation; types of companies] *§ 1. This Act regulates the creation, organization, functioning, dissolution, merger, division and transformation of commercial companies.*

§ 2. Commercial companies shall include: a registered partnership, […] a limited liability company and a joint-stock company.

Art. 4. [Terms] *§ 1. The terms used in this Act shall mean: […]*

2) a capital company - a limited liability company and a joint-stock company,

3) a single-shareholder company – a capital company in which all the shares belong to one shareholder […].

Art. 22. [Definition; liability] *§ 1. A registered partnership is a partnership which operates a business under its own business name and is not another commercial company.*

Art. 151. [Purposes; situation of shareholders] *§ 1. A limited liability company may be incorporated by one or more persons for any purpose allowed by law, unless the law provides otherwise.*

§ 2. A limited liability company may not be formed solely by another single-shareholder limited liability company.

§ 3. The shareholders shall provide only the performances stipulated in the articles of association.

Art. 157. [Contents of articles; form] *§ 2. The articles of association shall be made in the form of a notarial deed.*

Art. 301. [Promoters; statutes; liability] *§ 1. A joint-stock company may be formed by one or more persons. A joint-stock company may not be formed exclusively by a single-shareholder limited liability company.*

§ 2. The statutes of the joint-stock company shall be made in the form of a notarial deed.

In the next subsection a conceptual model and implementation of the ontology on the basis of the above legal definitions will be shown.

3.2 Conceptualization and Implementation

The Methontology (described in [8] and [9]), an ontology construction methodology has been used as a base one in the building process with respect, however, to the exceptionality of legal ontologies in general. The ontology so far, defines more than 135 concepts and more then 35 properties and about 50 axioms.

Additional background of most fundamental structure of the ontology (properties hierarchy and symmetry; place in taxonomy of some basic concepts) may be found in [13] and therefore these information is omitted here. A basic sketch of the fragment of the architecture is placed on Fig. 1.

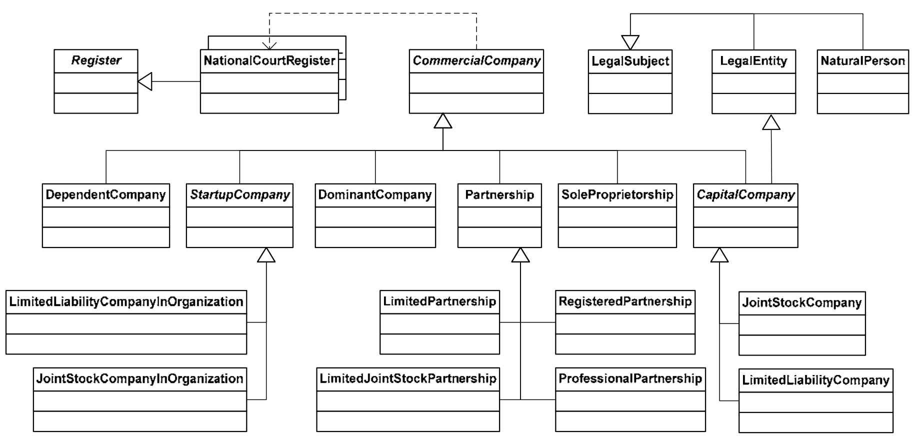


Fig. 1. Basic concepts of PCCC ontology

As we aim to model only selected parts of the legal domain we did not endeavor to ensure that all the domain knowledge is represented. The model is correct in terms of being consistent. A valuable estimation of both completeness and appropriateness is the one of the domain expert yet it is out of scope of this paper.

There is also the problem characteristic for every legal knowledge system connected with the fuzziness of reality compared with restricted language of law – however as we deal mostly with well-tailored legal definitions here in a well-qualified domain there is little space for uncertainty in this particular case.

4 Reasoning

Reasoners are used to perform a broad range of different tasks. They all are based on the knowledge defined in the knowledge base and use axioms and facts represented in a logical notation to produce certain output. The most common tasks include resolving problems of consistency checking, entailment, satisfiability and subsumption, all of which are provably decidable. Some reasoners also offer other functionalities i.e. explanation of their reasoning process, dedicated knowledge query languages, etc. A number of powerful reasoning systems are available for OWL using the DIG interface [10]. In our experiments we have used the RacerPro (Renamed ABox and Concept Expression Reasoner). RacerPro is a robust reasoning system, which support both OWL-Lite and OWL-DL documents, providing as well a highly optimized tableau calculus.

4.1 Case 1 – An Unlimited Partnership

An Unlimited Partnership is the example where two independent deduction paths leads to some conclusion about another entity. The case is significantly more complicated - especially in terms of complexity level and needed computing power – than the one reported in [13]. Another issue is that because of more distinct definition of PCCC of concepts involved in this case the logical formulae are considerably more meaningful.

The logical statements defining the concept of LegalEntity are presented below as (1) and (2). We used a standard logic formula notation in order to make the ideas standing behind the clauses more understandable. Both mentioned clauses define a necessary and sufficient conditions which corresponds to the equivalence relation in the logical notation whereas, for instance (8), represents the necessary condition statement in terms of Protégé terminology. The necessary and sufficient condition means that any instance that satisfies at least one of them is the instance of a class that defines the condition. A class that has at least one necessary and sufficient condition defined is thus complete as it can have instances that are not explicitly stated. The necessary condition defines minimal restriction put on any instance of the class. One cannot predict whether the instance is an instance of the class with defined necessary condition unless it is explicitly stated. However such a statement allows giving negative premises that exclude the instance from the set of given class's objects. The third (3) clause is a general restriction on defined concepts.

$$\forall x, y: \text{LegalEntity}(x) \wedge \text{NationalCourtRegisterA}(y) \wedge \text{isRegisteredIn}(x, y) \Leftrightarrow \text{CapitalCompany}(x) . \tag{1}$$

$$\forall x, y: \text{LegalSubject}(x) \wedge \text{CompanyAgreement}(y) \wedge \text{isEstabilishedOn}(x, y) \Leftrightarrow \text{CommercialCompany}(x) . \tag{2}$$

As a background for this example we will introduce some new basic facts from our knowledge model. We define RegisteredPartnership as in (3).

$$\forall x, y: \text{Partnership}(x) \wedge \text{RegisteredPartnershipAgreement}(y) \wedge \text{isEstabilishedOn}(x, y) \Leftrightarrow \text{RegisteredPartnership}(x) . \tag{3}$$

We add necessary and sufficient axioms for the superclass of RegisteredPartnership – namely the Partnership as stated in (4).

$$\forall x: \text{CommercialCompany}(x) \wedge \neg (\text{CapitalCompany}(x)) \wedge (\text{isRegisteredIn} = 1) \Leftrightarrow \text{Partnership } (x) . \tag{4}$$

Now some declarative axioms mentioning LegalAgreements and their respective subclasses will be brought in. The concept hierarchy is depicted on Fig. 2.

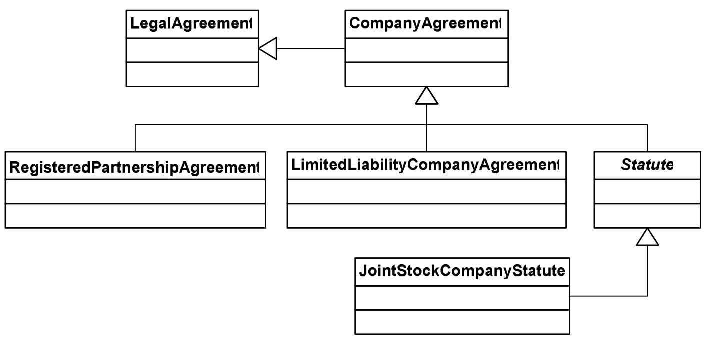


Fig. 2. UML class diagram of legal agreements hierarchy

One can see that we derive all specific types of CompanyAgreements from the civilistic notion of concept of LegalAgreement which is the term characteristic rather for broader domains such as civil law or even law theory. Nonetheless, subtypes which are most important in the commercial law domain are (described rather than) defined in Polish Commercial Companies Code. Thus, some of our restriction imposed on this terms are enlisted below. In (5) equivalence has been introduced in order to give sufficient conditions for JointStockCompanyStatute.

$$\forall x, y: \text{Statute}(x) \wedge [\text{JointStockCompany } (y) \vee \text{StartUpJointStockCompany } (y)] \wedge \text{isEstabilishing}(x, y) \Leftrightarrow \text{JointStockCompanyStatute}(x) . \tag{5}$$

Whereas (6) together with the above taxonomy defines the concept of RegisteredPartnershipAgreement.

$$\forall x, y\text{: CompanyAgreement }(x) \wedge \text{WrittenForm}(y) \wedge \text{hasForm}(x, y) \Leftrightarrow \text{RegisteredPartnershipAgreement}(x)\ . \tag{6}$$

$$\forall x, y\text{: RegisteredPartnership}(y) \wedge \text{isEstabilishing}(x, y) \Leftrightarrow \text{RegisteredPartnershipAgreement }(x)\ . \tag{7}$$

$$\forall x, y\text{: } \neg\ [\text{RegisteredPartnershipAgreement }(x) \wedge \text{OralForm}(y) \wedge \text{hasForm}(x, y)]\ . \tag{8}$$

Actually, (6) reflects the fact that in the registered partnership's case the form of the agreement is its distinctive feature – as wrote down in:

Art. 23. [Form of articles] *The articles of association shall be made in writing, or else they shall be invalid.*

Effectively, this means that no "stronger" formal form of the agreement is needed.

Now let us look at the possible reasoning pattern. Let us assume that we define a number of instances in our knowledge base, namely aCompany of type CommercialCompany and someCompanyAgreement as CompanyAgreement. The initial attributes are filled with values according to those presented on Fig 3.

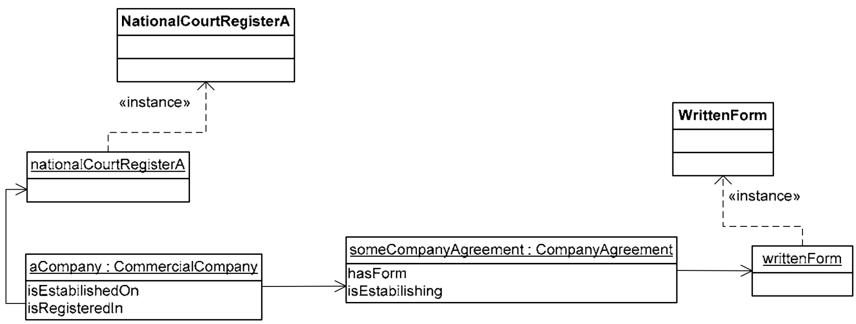


Fig. 3. UML diagram of instances demonstrating the initial acquired knowledge for RegisteredPartnership case

From (6) we induce that someCompanyAgreement is in fact RegisteredPartnershipAgreement supposing that (11) is false. However a common sense suggests that the latter must be false as an agreement cannot have more than one form (i.e. it cannot be legally binding at the same time referring to regulation of the same facts within the same legal relationship; although some parties may negotiate / arrange the terms and then sign a contract or change orally the terms after the written agreement has been signed). Later, we will show how to ensure such a constraint takes effect. Once we conclude that we deal with RegisteredPartnershipAgreement we may move on to the company itself.

Since (4) infers that aCompany is a Partnership (per definition it is CommercialCompany) and implicitly we know that it is not a CapitalCompany (however we do not have such knowledge stated explicitly – it may be proved on the basis of definition of CapitalCompany). The isRegisteredIn clause in (4) is also in concord with the instance definition. Next by virtue of (3) combined with facts that:

- someCompanyAgreement is identified as RegisteredPartnershipAgreement.
- aCompany has been proved to be a Partnership.

We come with the conclusion that indeed the aCompany is an instance of class RegisteredPartnership with all imaginable consequences of this fact. For instance our system may next draw some conclusions on the financial responsibility of partners involved in this type of partnership or on rights (legal capabilities) and duties of the parties imposed by law in relation to the partnership itself – we will show some possibilities in the third case.

Now let us come back to the point of deduction process about someCompanyAgreement. We may extend the whole case by adding other axioms as shown in (9) thru (11).

$$\forall x, y\text{: CommercialCompany}\ (y) \wedge \text{isEstabilishing}\ (x, y) \Leftrightarrow \text{CompanyAgreement}\ (x)\ . \tag{9}$$

$$\forall x\text{: } \neg\ \{\text{CompanyAgreement}\ (x) \wedge \neg\ [\text{LegalAgreement}\ (x)]\}\ . \tag{10}$$

$$\forall x\text{: hasForm} \neq 1 \Rightarrow \neg\ [\text{LegalAgreement}\ (x)]\ . \tag{11}$$

The latter implication is the formally described constraint that we mentioned earlier on the occasion of consideration whether an agreement can have more than one form simultaneously. The (9) and (10) statements can be viewed together. The first one imposes a necessary and sufficient condition for identifying CompanyAgreement whereas the other one introduce a must-be restriction for the same class of objects. The second one may be understood as “there are no agreements that bring any company to life that are not legal agreements” and it is attached here only for explanation reasons as generally it can be derived from the ontological taxonomy.

With these additional rules included into our reasoning system a more powerful and comprehensive inference is applicable. Now we do not need to specify the type of the contract erecting the company only the legal consequences (or a mere nature) of the agreement has to be specified to draw the same conclusions as before. In fact by adding the three axioms we enabled the inference engine to perform much more general knowledge processing by stepping out from the domain of PCCC into civil code domain.

4.2 Case 2 – Reductio Ad Absurdum and a Limited Liability Company

In the second case - for transparency - we assume that all model excerpts form previous example actually hold. What we are going to show now is the use of logical law of non-contradiction together with some elements of our Commercial Code model in order to finally come to an interesting conclusion about few false premises included in the prepared knowledge base. The classical form of the law of non-contradiction in the formalism of propositional logic might be noted as follows:

$$\forall \pi: \neg [\pi \wedge \neg (\pi)] . \tag{12}$$

Which statement in fact holds very common sense truth that the same statement cannot be "true" and "false" at the same time and under the same circumstances.

From [13] we recall the exemplary axioms defining the JointStockCompany:

$$\forall x, y: \text{CommercialCompany}(x) \wedge \text{JointStockCompanyStatute}(y) \wedge \text{isEstabilishedOn}(x, y) \Leftrightarrow \text{JointStockCompany}(x) . \tag{13}$$

$$\forall x, y: \neg \{\text{JointStockCompany}(x) \wedge \text{JointStockCompanyStatute}(y) \wedge \neg [\text{isEstabilishedOn}(x, y)]\} . \tag{14}$$

The conversion of the legal text's excerpts may be as demonstrated below. The (15) allows identifying the instances of class SoleProprietorship which are all those CommercialCompanies that has only one owner.

$$\forall x: \text{CommercialCompany}(x) \wedge (\text{isObjectOfProprietaryRight} = 1) \Leftrightarrow \text{SoleProprietorship}(x) . \tag{15}$$

The subsequent axioms (16) thru (19) define a LimitedLiabilityCompany. The definition is similar to the one of JointStockCompany however with some exceptions reflecting the differences in regulations between those two types of capital companies.

The UML diagram represents respectively: the hierarchy of the SingleShareholderLimitedLiabilityCompany which multiple-inherits from SoleProprietorship and LimitedLiabilityCompany – concepts defined below.

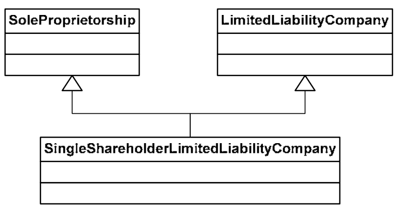


Fig. 4. UML class diagram of single-shareholder limited liability company hierarchy

$$\forall x, y: \text{CapitalCompany}(x) \wedge \text{LimitedLiabilityCompanyAgreement}(y) \wedge \text{isEstabilishedOn}(x, y) \Leftrightarrow \text{LimitedLiabilityCompany}(x) . \tag{16}$$

$$\forall x, y: \text{CapitalCompany}(x) \wedge \text{ShareholdersMeeting}(y) \wedge \text{hasProprietaryBody}(x, y) \Leftrightarrow \text{LimitedLiabilityCompany}(x) . \tag{17}$$

$$\forall x, y: \neg \{(\text{isObjectOfProprietaryRight} \geq 2) \vee [\text{SingleShareholderLimitedLiabilityCompany}(y) \wedge \text{isObjectOfProprietaryRight}(x, y)]\} \Rightarrow \neg (\text{LimitedLiabilityCompany}(x)) . \tag{18}$$

$$\forall x, y: \neg \{\text{LimitedLiabilityCompany}(x) \wedge \text{LimitedLiabilityCompanyAgreement}(y) \wedge \neg [\text{isEstabilishedOn}(x, y)]\} . \tag{19}$$

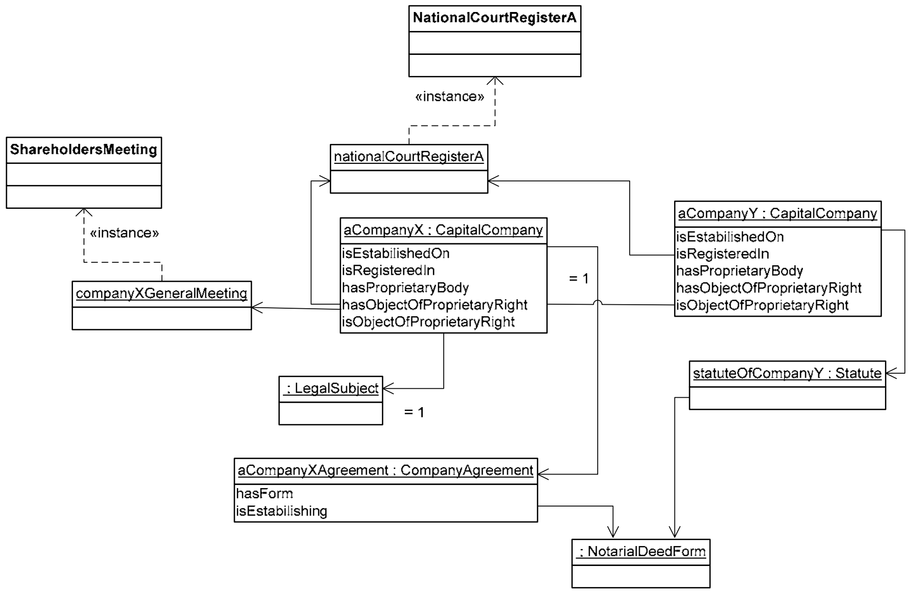


Fig. 5. The assumed initial state of knowledge for second example

We assume that the initial state of our knowledge is given and it is as declared on Fig. 5.

As shown on the above figure, for the sake of this case two companies are given, namely aCompanyX and aCompanyY. Some details are told as a basic characteristic of those instances. We know that aCompanyX is a sole possessor of aCompanyY and has only a single possessor. What is more, we also know that aCompanyY is a CapitalCompany. Still there are a number of additional facts which may be further derived in order to extend our knowledge.

We may simulate the reasoning steps performed originally by the inference engine and because of simplicity reasons we will divide this process into three phases. The first one includes:

- Basing on the initial knowledge and the definition contained in (16) thru (17) we may show that aCompanyX is a LimitedLiabilityCompany (it is a CapitalCompany and it hasProprietaryBody of ShareholdersMeeting type).
- The negative premises as declared in axioms (18) and (19) do not hold.
- Thus, we may infer that the aCompanyX is a LimitedLiabilityCompany.
- Although it is now obvious to the reader the reasoner need additional processing of the (16) to conclude that aCompanyX is a SoleProprietorship.
- All those above facts give reason to presume that the aCompanyX is a SingleShareholderLimitedLiabilityCompany because [(aCompanyX $\supseteq$ LimitedLiabilityCompany) $\wedge$ (aCompanyX $\supseteq$ SoleProprietorship)].

  From the other point of view (the second phase) we disprove:

- Using axioms ranging from (13)-(14) and (1)-(2) a reasoning process similar to the one from the case in [13] may be conducted with the result that aCompanyY is a JointStockCompany.
- From (15) we presume that aCompanyY is also a SoleProprietorship.
- We know that one and the only possessor of aCompanyY is aCompanyX.
- However (18) for LimitedLiabilityCompany exclude such situation (we assume that the exclusion works respectively for JointStockCompany). Finally, we come to the conclusion that the aCompanyX is not a LimitedLiabilityCompany.

The third and the last phase consists of taking the above outcomes of two initial phases and insert them into the logical formula of the law of non-contradiction. It is evident that the "π" clause in our case reads: "xCompany is a LimitedLiabilityCompany". After such substitution in (12) the statement is of course true, nonetheless we are aware of the ridiculous situation of absurd indicating that at least some of the early premises of the inference process were apparently false (we may safely assume that the axioms are well-formed). What does it mean in practice?

In this case there are two sources of the conflict. The one is the fact that aCompanyX has a body named GeneralMeeting which on the basis of PCCC implies that it is a LimitedLiabilityCompany and in addition it is a single-owner company as far as we are aware of it. On the other hand aCompanyX is a sole proprietor of aComapnyY – a single-shareholder capital company. After the reasoning we proved that at least one of these facts cannot be true. For instance: if we defy the piece of evidence about the body the rest of facts become coherent although in such case one cannot say anything about the type of aCompanyX – except that it is not a PLC. Another acceptable set of facts remains after excluding the information about the cardinality of the set of owners of (at least one of) respective companies. The coherence in this case is connected to not holding of axioms (2) and (19). An interesting point of view on this issue is laid in [11]. The author points out that "[…] a source of defeasibility is conflicting rules. Two rules conflict if one of them is used to derive P and another (not P). To resolve these conflicts, we need to be able to reason (i.e. argue) about which rule has a priority. To support reasoning on rule priorities, the rule language includes a built-in predicate over rules, prior, where (prior r^1 r^2) means that rule r^1 has priority over rule r^2".

5 Conclusion

This paper has introduced few examples of advanced legal knowledge modelling in the domain of Commercial Law. Two cases demonstrated how the knowledge model together with additional explicit knowledge in the form of facts may be used to perform reasoning tasks. The inference procedures are typically performed by dedicated inference engine. The conclusions reached in this way may be used in a number of applications. In addition to that there exist some representation difficulties when developing advanced legal models such as the one presented.

As a future work we intend to recognize the use of more advanced reasoning system with the explanatory capabilities to produce results in context of NL output.

Secondly, we would like to work on expanding our model with new concepts but also with metamodeling structures. Adding a priority (meta)rule is a tempting vision of augmenting our legal model - further work have to be done to attain promising results.

References

1. Knublauch, H., Musen, M.A., Rector, A.L.: Editing Description Logic Ontologies with the Protégé OWL Plugin. In: International Workshop on Description Logics (DL 2004), Whistler, Canada (2004)
2. Despres, S., Szulman, S.: Construction of a Legal Ontology from a European Community Legislative Text, Legal Knowledge and Information Systems. In: Jurix 2004: The Seventeenth Annual Conference, Amsterdam, pp. 79–88 (2004)
3. Maedche, A., Staab, S.: Measuring Similarity between Ontologies. In: Gómez-Pérez, A., Benjamins, V.R. (eds.) EKAW 2002. LNCS (LNAI), vol. 2473, Springer, Heidelberg (2002)
4. Baader, F., et al.: The Description Logic Handbook. Cambridge University Press, Cambridge (2003)
5. Champin, P., Houben, G., Thiran, P.: Cross: an OWL wrapper for reasoning with relational databases. In: Parent, C., Schewe, K.-D., Storey, V.C., Thalheim, B. (eds.) ER 2007. LNCS, vol. 4801, Springer, Heidelberg (2007)
6. Euzenat, J., Valtchev, P.: Similarity-based ontology alignment in OWL-Lite. In: Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004), pp. 333–337 (2004)
7. Michałowska, K.: Polish commercial companies code: bilingual edition Polish-English / Eng. transl. by Katarzyna Michałowska, 3rd edn., Warszawa, C.H. Beck (January 2004)
8. Corcho, O., Fernández-López, M., Gómez-Pérez, A., López-Cima, A.: Building Legal Ontologies with METHONTOLOGY and WebODE. In: Benjamins, V.R., Casanovas, P., Breuker, J., Gangemi, A. (eds.) Law and the Semantic Web. LNCS (LNAI), vol. 3369, pp. 142–157. Springer, Heidelberg (2005)
9. Fernandez, M., Gomez-Perez, A., Juristo, N.: METHONTOLOGY: from Ontological Art towards Ontological Engineering. In: Proceedings of the AAAI 1997 Spring Symposium Series on Ontological Engineering, pp. 33–40 (1997)
10. Haarslev, V., Möller, R.: The DIG Description Logic Interface. In: Proceedings of the International Workshop on Description Logics (DL 2003), Rome, Italy, September 5-7 (2003)
11. Gordon, T.F.: Constructing arguments with a computational model of an argumentation scheme for legal rules: interpreting legal rules as reasoning policies. In: ICAIL 2007: Proceedings of the 11th international conference on Artificial intelligence and law, pp. 117–121. ACM Press, New York (2007)
12. Gangemi, A.: Design patterns for legal ontology construction. In: Proceedings of LOAIT 2007 II Workshop on Legal Ontologies and Artificial Intelligence Techniques (2007)
13. Abramowicz, W., Stolarski, P., Tomaszewski, T.: Reasoning in the Domain of Commercial Law Using Commercial Companies Code Ontology. In: Proceedings of JURIX 2007. Frontiers in Artificial Intelligence and Applications Series, IOS Press, Amsterdam (2007)
14. Benjamins, R., Casanovas, P., Breuker, J., Gangemi, A. (eds.): Law and the Semantic Web. LNCS (LNAI), vol. 3369. Springer, Heidelberg (2005)

A Supply Chain Management Approach to Logistics Ontologies in Information Systems

Joerg Leukel and Stefan Kirn

University of Hohenheim, Information Systems II, Schwerzstr. 35,
70599 Stuttgart, Germany
{joerg.leukel,kirn}@uni-hohenheim.de

Abstract. Logistics models in information systems describe activities, organizations, transportation means, goods, and services being involved in logistics processes. The problem with most current such models, however, is a lack of formal semantics which prevents automated data integration across organizational boundaries. In this paper, we take the perspective of supply chain management and employ a well-grounded model which provides core concepts of interorganizational logistics. The contribution is that we (1) propose referring to supply chain management for ontologizing logistics models and (2) provide definitions of core elements of logistics ontologies.

Keywords: Interorganizational Information Systems, Logistics, Ontologies, Semantic Integration, Supply Chain Management.

1 Introduction

Logistics concerns controlling and executing the flow of goods, services, and related information from sources to destinations. Logistics is a vertical function being important to almost any industry. It can be characterized by the involvement of multiple stakeholders in logistics processes and the need for coordination across organizational boundaries. Respective *logistics models in information systems* describe such processes and coordination mechanisms. For instance, ERP systems incorporate elaborated logistics data and process models, and allow for exchanging data via message exchange formats. The problem with such models is a *lack of formal semantics* which prevents automated data integration. This is in particular made complicated by the variety of respective models which is a good indication of the conceptual complexity and diversity of logistics.

In recent years, ontologies have attracted both industry and academia because of their potential contribution to solving integration problems in information systems. By providing to some extent consensual definitions of concepts and inter-relationships between these concepts in a domain of interest, ontologies represent a consolidated body of knowledge to which users can commit to. Current logistics ontologies, however, have not yet reached a high level of visibility and maturity. Surprisingly, few logistics ontologies are available on the Web, if any. Despite the relevance of the logistics domain, there is rather little interest by researchers.

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 95–105, 2008.

The objective of this paper is to contribute to the advancement of logistics ontologies. Unlike other attempts to building logistics ontologies by following an engineering approach including a requirements analysis, we take the *perspective of supply chain management* and reuse the existing body of knowledge contained in the SCOR model [1]. This model defines a comprehensive terminology and a set of semi-formal models of interorganizational logistics; due to its wide acceptance it can be regarded as a reference model. Though this model is not directly aimed at information systems, but at the design of supply chains. The contribution of our research is that we (1) propose referring to supply chain management for ontologizing logistics models and (2) provide definitions of core elements of logistics ontologies.

The present work contributes to a research framework which concerns *logistics systems under customization*. Logistics systems provide services which transform goods with regard to location, time, and quantity. The goal of this research framework is to make logistics massively customizable by means of information systems. Customization is a major trend that can be observed in many industries and markets. It says that customers ever more demand customized goods and services which are tailored to their specific needs [2]. Firms have to revise their strategies and operations to meet this challenge; management science has developed respective differentiation strategies, e.g., [3]. This trend essentially concerns logistics systems which can be seen as the backbone of any industry of tangible goods. A *customized logistics service* is one that is tailored to the specific needs of an individuale customer. The current work addresses the means how customers and suppliers specify and thus represent logistics systems in information systems. As such, we employ the SCOR model as a basis for describing activities in logistics, their properties, and inter-relations.

The remainder of this paper is as follows. Section 2 reviews related work. Section 3 introduces briefly the SCOR model. In section 4, we propose logistics ontologies being based on the SCOR model. In section 5, we outline use cases. Section 6 draws conclusions and points to future work.

2 Related Work

The related work can be grouped into two major areas: logistics ontologies and ontology engineering.

Logistics ontologies are a rather specialized subject which is being reflected in the number of work on such ontologies and available ontologies on the Web. For instance, both SchemaWeb[1] and the DAML Ontology Library[2] return only one entry each for ‘logistics’ and both are even no logistics, but manufacturing respectively product ontologies.

When widening the scope, one can identify the topic as part of other ontologies. Very often, these ontologies concern a particular domain or function within logistics. Next, we provide an overview: The work of Wendt et al. describes considerations on how to derive common logistics concepts for scheduling from merging two domain-specific ontologies [4]; however, the planned ontology has not been published. Pawlaszczyk et al. describe the role of logistics ontologies in mass customization and

---

[1] http://www.schemaweb.info
[2] http://www.daml.org/ontologies

consider the Enterprise Ontology [5] as a starting point without giving a specification [6]. Haugen and McCarthy propose to extend the REA Ontology which concerns internal accounting to supporting logistics and e-commerce [7]; Gailly and Poels provide a methodology for defining this ontology using UML and OWL [8]. All these works aim at providing some basic concepts of particular logistics systems and thus the resulting ontologies remain quite shallow compared to the true complexity of logistics systems. In particular, such ontologies do not provide a sufficient set of concepts and inter-relations for supporting customization which goes beyond generic logistics.

Hofreiter and Huemer show how to derive RDF ontologies from UML-based data exchange specifications [9]; though this work is not confined to exchanging logistics data. Fayez et al. propose to use an OWL representation of the SCOR model for supply chain simulation, though they do not provide details on their implementation [10].

Brock et al. argue against the use of logistics ontologies because of the 'rigid and inflexible' nature of ontologies which would contradict characteristics of logistics [11]. In particular, they claim that it would be unrealistic to believe in formulating an 'all-inclusive canon that would stand the test of time'; Brock et al. relate this proposition to information systems in general, and propose to define rather light-weight abstractions such as multi-perspective taxonomies for the logistics domain.

Table 1 summarizes key properties of the related work on logistics ontologies.

Table 1. Related Work on Logistics Ontologies

Author	Domain or Function	Language	Specification
[4]	Scheduling; Manufacturing, Healthcare	Ontolingua	Not available
[6]	Mass Customization	Ontolingua	Not available
[7]	E-Commerce and Supply Chains	None	Not available
[8]	E-Commerce and Supply Chains	UML, OWL	Not available
[9]	Data Exchange	RDF	Not available
[10]	Supply Chain Simulation	OWL	Not available
[11]	Business Logistics	n/a	n/a

To the best of our knowledge, there exist two approaches that link the SCOR model to ontologies: Fayez et al. consider SCOR as part of a wider multi-view ontology and do not provide details on their implementation [10]. Another ontology-based version under the label SCOR+ has been marketed by the firm Productivity Apex; though this is a proprietary approach and no further information is available to the public [12].

Ontology engineering aims at providing methods, languages, and tools for developing new ontologies and maintaining existing ones. An overview of this area can be found, for instance, in [13]. For the purpose of our work, we distinguish two approaches which start from different *prerequisites*.

(a) Building new ontologies: This approach requires a systematic engineering process which includes, among others, a detailed requirements analysis and definition by, for instance, involving end-users and/or referring to relevant theories and models. Adopting this approach thus calls for respective analysis of logistics application scenarios, theories and models.

(b) Ontologizing existing models: This approach takes an existing model, specification, or standard and raises the degree of formal semantics by employing a respective ontology language ("ontologizing"). It necessarily changes the language used, but leaves most of the original model unchanged and thus reuses knowledge contained in the original. Literature yields a rich set of methods for model conversion in general (e.g., from XML to RDF; from UML to OWL) and also domain-specific adoptions of such methods (e.g., [14] on EDI, [15] on e-catalog data).

Considering the review of current research on logistics ontologies in the former part of this section, we have to state that approach (a) has so far attracted few researchers. In order to reuse existing knowledge relevant to logistics to a greater extend, we follow the direction of (b). Though, we do acknowledge that the former direction opens prospective and fruitful avenues of research.

3 Supply Chain Operations Reference Model

In this section, we briefly introduce the Supply Chain Operations Model (SCOR model) [1]. A supply chain is a system of entities being involved in producing, transforming and/or moving a good or service from suppliers to customers. SCOR provides a comprehensive set of means for modeling supply chains. Unlike generic process modeling languages, it defines a huge number of domain-specific elements for distinguishing different means of manufacturing and moving goods. SCOR has been developed by The Supply-Chain Council (SCC), an independent not-for-profit firm with more than 1,000 corporate members. It was first introduced in 1996 and is currently available in version 8.0.

The documentation of SCOR comprises of 540+ pages; it is available to the SCC of which we are member. The general structure and approach are also described in public documents which can be obtained from the SCC's website. All SCOR elements are defined in natural language and semi-formally (tables and figures). Interrelations between elements are partly defined by referring to identifiers (i.e., metrics and process elements). Metrics are described verbally and, where possible, formally.

SCOR consists of a model stack of top level, configuration level, and process element level as follows:

Top level: This level distinguishes five core management processes called 'process types' that are relevant for all firms in a supply chain. These are: plan, source, make, deliver, and return.

Configuration level: It provides for each process type of the top level a set of 'process categories' which represent different operational strategies that a company pursues. For instance, the process categories for 'source' represent sourcing strategies. By connecting process categories, a company can describe its logistics processes in a so called 'process map'. In addition, metrics and best practices are assigned to categories.

Process element level: It decomposes the process categories by adding (1) process element definitions and (2) process element information inputs / outputs. For instance, a particular source category may be decomposed into process elements for receiving, verifying, and finally stocking the good. Metrics and best practices of the former level appear here also in greater detail. This level provides the most comprehensive set of modeling primitives. SCOR defines 295 information entities which can be input or output of process elements. They range from atomic entities such as ‘vendor lead time’ to complex and consolidated ones such as ‘payment terms’ and ‘service levels’.

4 Proposal of Logistics Ontologies

In this section, we propose a set of logistics ontologies being based on the SCOR model. We specify the ontologies in OWL and use a customized graphical notation which covers the following language constructs: Class, Object Property, Datatype Property, and subClassOf.

4.1 Logistics Top Level Ontology

The purpose of the Logistics Top Level Ontology is to define the scope of the ontology. On the respective level, SCOR distinguishes five process types which thus are sub classes of the generic process type class. Figure 1 depicts the OWL representation.

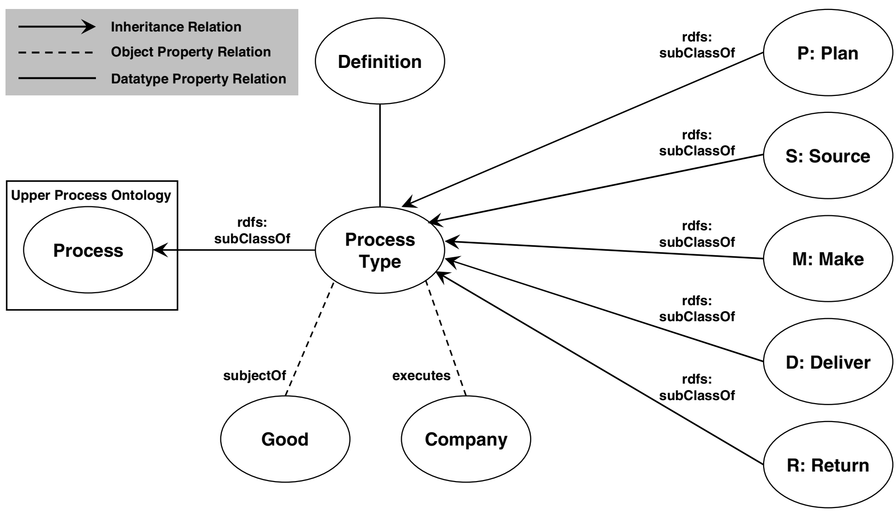


Fig. 1. Logistics Top Level Ontology

When reconstructing the semantics of SCOR by studying the documentation, we identify two relationships to other concepts which are not described explicitly. First, goods are subject of logistics processes. Second, logistics processes are executed by companies. In both cases, SCOR does not limit the cardinality of these relationships which also holds for logistics in general.

The only other information that can be taken from SCOR is a natural language definition of each process type; this information is stored in a respective Datatype Property.

Since a set of logistics processes describes a supply chain, there are constraints on how to interconnect processes depending on the process type. For instance, 'source'→'make'→'deliver' reflects the flow of goods to customers while 'return' is in the opposite direction. Specifying such constraints in the ontology requires means for expressing *preconditions* of a process type.

Here, we employ an externally defined Upper Process Ontology which provides respective classes and interrelations; hence we avoid defining a custom ontology. We choose OWL-S which contains a process model for describing web services[3].

The following OWL statement links to the respective 'Process' class of OWL-S:

```
<rdf:RDF [...] xmlns:process="http://www.daml.org/services/owl-s/1.1/Process.owl#">
[...]
<owl:imports>
  <owl:Ontology rdf:about="&process;"/>
</owl:imports>
[...]
<owl:Class rdf:ID="ProcessType">
  <rdfs:subClassOf rdf:resource="#Process"/>
</owl:Class>
```

4.2 Logistics Process Type Ontology

The Logistics Process Type Ontology provides definitions of operational strategies underlying a process type. Basically, there are three strategies that affect all process types and describe whether the good is (1) on stock, (2) made-to-order thus manufactured for a specific customer order, or (3) engineered-to-order thus it is designed and manufactured specific to a particular customer requirement. In SCOR, this distinction is justified by respective strategies of how manufacturers meet customer demand. These basic strategies also serve for describing interorganizational logistics, since they relate to sourcing and delivery in SCOR.

Building a respective ontology has to decide whether each process category should be represented by a class or instance of a class. The latter approach, however, would prevent defining process category constraints by the mechanism provided in the Upper Process Ontology; hence we choose the former.

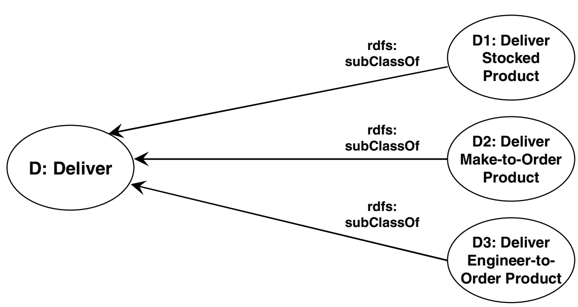


Fig. 2. Logistics Process Type Ontology for 'D: Deliver'

[3] http://www.daml.org/services/owl-s

Figure 2 shows the ontology for 'deliver'; those for 'source' and 'make' look very much the same while the 'return' ontology includes subclasses for returning defective, excessive, and other types of goods.

4.3 Logistics Process Category Ontology

The Logistics Process Category Ontology provides not only a more detailed level of logistics processes by means of process elements, but also enriches process categories with metrics and best practices. Process elements introduce the lowest level of abstraction by specializing process categories.

Metrics allow assessing the performance of a process category and best practices describe empirically proofed means for achieving good performance. In the ontology, both are modeled by Object Properties. Figure 3 does not show the list of metrics and best practices for the category 'Deliver Stocked Product' directly. The reason is that we subsume the entire set of metrics and best practices in separate parts. We then define the allowed metrics and best practices for each category by a constraint on the 'measures' and 'supports' relation.

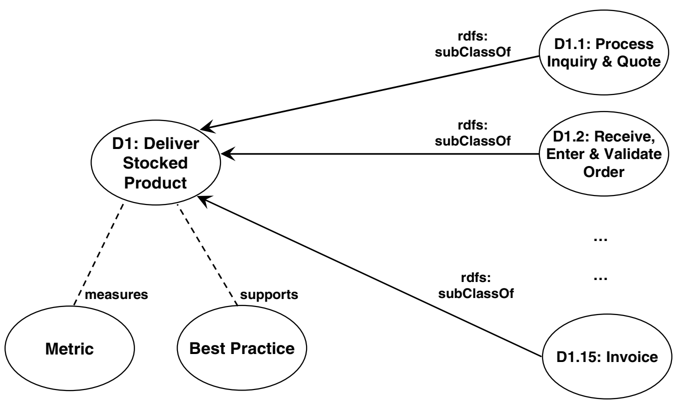


Fig. 3. Logistics Process Category Ontology for 'D1: Deliver Stocked Product'

Figure 4 shows a cutout of the Metrics Ontology which showcases the hierarchy of metrics as well as Datatype Properties for definition and calculation. The respective ontology for best practices arranges all best practices as subclasses and includes one Datatype Property 'definition' only (not shown here due to space limitations).

4.4 Logistics Process Element Ontology

The Logistics Process Element Ontology is rather small, since most interrelated classes and properties of its process elements classes are being inherited from its parent classes, i.e., definition, good, company, allowed sequences, metrics, and best practices.

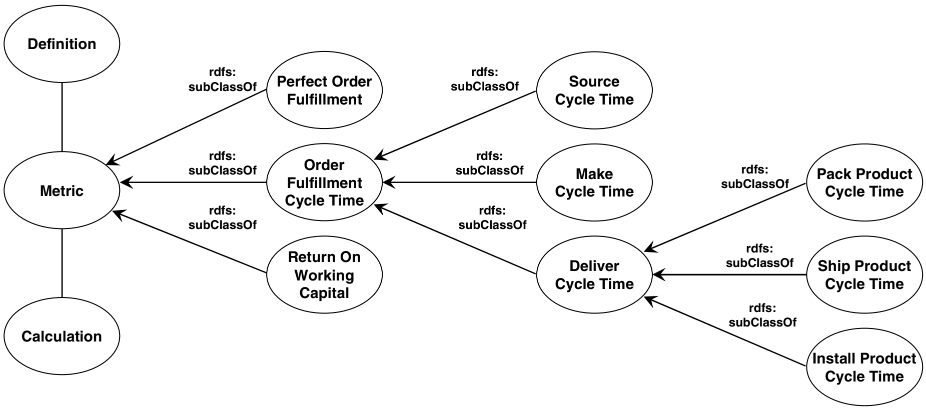


Fig. 4. Metrics Ontology (Cutout)

This level adds input information required for executing a process element and output information as the result of it. The respective ontology for a particular process element is shown in Figure 5.

Fig. 5. Logistics Process Element Ontology for 'D1.2: Receive, Enter & Validate Order'

Similarly to the case of metrics and best practices, Figure 5 does not list the inputs and outputs, since these are defined in another part of the ontology which provides the entire collection of information classes. Again, the allowed information is defined by a constraint on the two Object Property relations. The Information Ontology does not distinguish input and output but defines information only (see Figure 6).

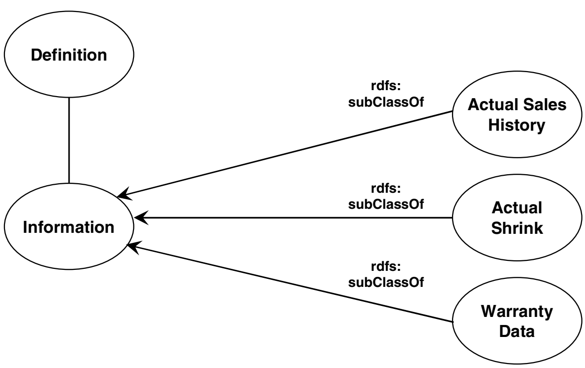


Fig. 6. Information Ontology (cutout)

5 Use Cases

In this section, we outline two use cases of the proposed logistics ontologies. Both cases are being based on the annotation of instance data. Thus we assume that the ontologies serve as consensual definitions to which users commit. For example, current means of data storage in ERP systems and interorganizational data exchange could make use of ontologies by annotating these instances accordingly, i.e., data records and messages. Tagging data, therefore, limits the required effort to changing existing data management means and processes.

(a) Searching for and aggregating logistics processes: In this use case, individual instances or the space of all instances are queried for particular processes according to the SCOR levels. While instances will include references to the lowest level of process elements, one can determine the respective super-processes by following the subclassOf relations, i.e., determine process categories and process types. This procedure allows aggregating diverse process instances on the lowest level to categories and types and thus presents a high-level picture of the current process space.

(b) Reconstructing the logistics network: In this use case, the process space is queried for a particular good and/or company in order to gain information about *all* associated process instances. The rationale is that single instance data does not provide information how the respective instance – and the associated good and company – are related with other instances and thus other goods and companies. By querying for a particular good, one can retrieve all process instances that relate to the good. Due to the transitive nature of the subclassOf relations in the process hierarchy as well as of the Object Property relations, one can reconstruct those parts of the logistics network which are relevant to the particular good; thus it allows viewing the process space form the perspective of a good.

6 Conclusions

This paper aimed at advancing the state of logistics ontologies in information systems. By taking the perspective of supply chain management, we were able to reuse existing knowledge contained in the SCOR model and applied an ontology language to 'ontologize' selected parts of this model. Therefore, we make the following contributions: we (1) propose referring to supply chain management for ontologizing logistics models and (2) we provide definitions of core elements of logistics ontologies.

The implications of our research are two-fold: First, few logistics ontologies are currently available and they lack comprehensiveness and domain coverage; thus, the proposed core elements help filling this gap. Second, a formal representation of the SCOR model could also contribute to its usage for information systems design, since it allows an easier adoption by accessing a machine-readable representation.

This research can be seen as initial steps required for enabling a semantic description of the logistics domain. In particular, such description means are required for logistics systems under customization. Customization materializes in logistics in the concept of customized logistics services. Such a service distinguishes from a standardized or off-the-shelf service in the degree of pre-specification, thus which characteristics of the service are specified in advance by the provider and which characteristics can be determined by the customer. By defining the core concepts of logistics systems and detailling such concepts by means of taxonomies for processes and metrics, we provide a basis for a rich description by providers and customers.

Our proposal has several limitations which also point to additional work required to arriving at truly comprehensive logistics ontologies. First, we base our ontologies on the SCOR model which does not provide formal semantics, thus we had to reconstruct the intended semantics by manually studying its documentation. The process of ontologizing is not deterministic and involves choosing alternative ways of modeling. Second, the current ontologies do not reflect the semantics of the entire SCOR model (due to the size of the original model); therefore we have focused the core concepts. Third, we did not specify the ontologies to the full extent (e.g., constraints). Fourth, the SCOR model does not aim at covering interorganizational logistics to the full extend, thus the ontologies lack some characteristic concepts of this domain. This observation is in particular true for transformations of goods which are not explicitly modeled; there are only processes such as ‘pack’ and ‘deliver’ that do not describe formally modifications of goods in terms of quantity, packing, place, and time.

Despite these shortcomings, we believe that the presented approach and the current ontologies provide a sound base for extending the ontologies in the directions outlined above.

Acknowledgement

This work has been supported by the BREIN project (http://www.gridsforbusiness.eu) and has been partly funded by the European Commission’s IST activity of the 6th Framework Programme under contract number 034556. This paper expresses the opinions of the authors and not necessarily those of the European Commission. The European Commission is not liable for any use that may be made of the information contained in this paper.

References

1. Supply-Chain Council: Supply Chain Operations Reference Model (SCOR®) Version 8.0 (2006), http://www.supply-chain.org
2. Coates, J.: Customization promises sharp competitive edge. Research in Technology Management 38, 6–7 (1995)
3. Spring, M., Dalrymple, J.F.: Product customization and manufacturing strategy. International Journal of Operations & Production Management 20, 441–467 (2000)
4. Wendt, O., Stockheim, T., Grolik, S., Schwind, M.: Distributed Ontology Management Prospects and Pitfalls on Our Way Towards a Web of Ontologies. In: Dagstuhl Workshop (Event Nr. 02212) - DFG-SPP 1083 (2002)
5. Uschold, M., King, M., Moralee, S., Zorgios, Y.: The Enterprise Ontology. Knowledge Engineering Review 13, 32–89 (1998)

6. Pawlaszczyk, D., Dietrich, A.J., Timm, I.J., Otto, S., Kirn, S.: Ontologies Supporting Cooperation in Mass Customization – A Pragmatic Approach. In: International Conference on Mass Customization and Personalization – Theory and Practice in Central Europe (2004)
7. Haugen, R., McCarthy, W.E.: REA, a semantic model for Internet supply chain collaboration. In: OOPSALA 2000 Business Objects and Component Design and Implementation Workshop VI: Enterprise Application Integration (2000)
8. Gailly, F., Poels, G.: Development of a formal REA-ontology representation. In: Open INTEROP Workshop on Enterprise Modeling and Ontologies for Interoperability (EMOI), CEUR Workshop Proceedings, Porto, Portugal, vol. 160 (2005)
9. Hofreiter, B., Huemer, C.: B2B Integration – Aligning ebXML and Ontology Approaches. In: Shafazand, H., Tjoa, A.M. (eds.) EurAsia-ICT 2002. LNCS, vol. 2510, pp. 339–349. Springer, Heidelberg (2002)
10. Fayez, F., Rabelo, L., Mollaghasemi, M.: Ontologies for Supply Chain Simulation Modeling. In: Kuhl, M.E., Steiger, N.M., Armstrong, F.B., Joines, J.A. (eds.) 2005 Winter Simulation Conference, pp. 2364–2370 (2005)
11. Brock, D.L., Schuster, E.W., Allen, S.J., Kar, P.: An Introduction to Semantic Modeling for Logistical Systems. Journal of Business Logistics 26, 97–118 (2005)
12. Productivity Apex: Ontology-Based Supply Chain Process Mapping (SCOR+) (2008), http://www.productivityapex.com/products/scor.asp
13. Fernández-López, M., Gómez-Pérez, A.: Overview and analysis of methodologies for building ontologies. Knowledge Engineering Review 17, 129–156 (2002)
14. Foxvog, D., Bussler, C.: Ontologizing EDI: First Steps and Initial Experience. In: IEEE International Workshop on Data Engineering Issues in E-Commerce (DEEC 2005), pp. 49–58. IEEE Computer Society, Los Alamitos (2005)
15. Leukel, J., Hepp, M., Schmitz, V., Tribowski, C.: Ontologizing B2B Message Specifications: Experiences from Adopting the PLIB Ontology for Commercial Product Data. In: IEEE International Conference on e-Business Engineering (ICEBE 2006), pp. 146–153. IEEE Computer Society, Los Alamitos (2006)

Transforming Existing Knowledge Models to Information Extraction Ontologies

Marek Nekvasil, Vojtěch Svátek, and Martin Labský

Department of Information and Knowledge Engineering, University of Economics, Prague, Winston Churchill Sq. 4, 130 67, Prague 3, Czech Republic
{nekvasim,svatek,labsky}@vse.cz

Abstract. Various knowledge models are widely adopted nowadays and many areas are taking advantage of their existence. On one hand there are generic models, domain ontologies that are used in fields like AI and computer knowledge-aware systems in general; on the other hand there are very specific models that only come in use in very specific areas like software engineering or business analysis. In the domain of information extraction, so-called extraction ontologies are used to extract and semantically annotate data. The aim of this paper is to propose a method of authoring extraction ontologies by reusing other pre-existing knowledge models. Our priority is maintaining the consistence between the extracted data and the existing models.

Keywords: information extraction, ontology, UML, business models.

1 Introduction

Every model can basically be looked upon as an abstraction of reality according to a certain conceptualization. If the model can be expressed as a formal specification we can call it *ontology* according to the original T. Gruber's definition [4]. Once a model is represented as a concrete artifact, it can support communication, analysis and elaboration of the relevant aspects of the underlying domain.

It is now often assumed that the use of ontologies can bring the required flexibility to many disciplines, and we believe one of them is information extraction; and *web information extraction* (WIE) in particular. Hot WIE applications include structured searches over multiple 3rd party websites (e.g. product searches over many e-shops); finding the best offers in e-auctions, web page content quality assessment, question answering etc. In the field of WIE it is possible to distinguish several trends in the last few years. The *wrapper-based* approach is widely adopted in today's business spheres; it is based on structural information in the HTML documents. Although it is quite reliable, it is not only domain dependent but moreover document-structure dependent and thus individual extraction tasks are not very reusable. The second, *inductive*, approach is built upon statistical learning and/or language processing. The drawback of this approach is that it requires large corpora of annotated data as a base for the learning. In addition, both of these approaches usually provide the extracted

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 106–117, 2008.

data in a form that is not enough semantically structured for further use in knowledge-based systems.

Consequently, a third approach was formed, with focus on semantic annotation of extracted data, namely with a tendency for pushing structured ontologies towards the actual extraction process, in the role of extraction models, which can be referred to as *extraction ontologies* [2] when properly formalized. An extraction ontology is typically designed by a human with a specific extraction task in mind. For the actual extraction it may utilize multiple sources of extraction knowledge: hand-crafted extraction patterns, machine-learning classifiers trained using available training data and formatting regularities found in the analyzed documents. The possibility to quickly achieve a reasonably functional prototype application even when only one of the extraction knowledge sources is available is a crucial benefit of this approach.

We think that a strict single-purpose hand-crafting of such extraction ontologies from scratch is tedious because it is very demanding to author such an ontology manually (it is often done in iterations). Moreover it can introduce inconsistencies in relation to other business models and knowledge-bases but the mutual consistency is eligible in both academic spheres and enterprise environment. In this paper we hypothesize that extraction models can be crafted via reuse of existing meta-models that are already present in the company or freely available on the internet in ontology libraries. This reuse should improve further processing of any data annotated (or extracted) using the extraction ontology in terms of other knowledge models and hopefully even lower the costs of its creation because of lessening the need of thorough prior domain-analysis.

The paper is structured as follows. The next section describes the nature of extraction ontologies with emphasis on its relation to other kinds of knowledge models. After that follow three sections that dissect the possibilities of reusing various sources for the construction of extraction ontologies; firstly the ordinary domain ontologies are taken into account, after that we consider the potential of knowledge stored in UML diagrams and finally we focus on reusing other models very commonly used in industry, namely business process models and relational models.

2 Presentation Ontologies

Extraction ontologies define the concepts the instances of which are to be extracted from the documents in terms of WIE, in the sense of various attributes, their allowed values as well as higher level constraints (such as e.g. cardinality). Following the terminology coined in [6], an extraction ontology can be systematically viewed both as an information ontology and a knowledge ontology, depending on its actual content. It is possible to spot a few layers in the structure of an extraction ontology:

1. An extraction ontology contains concepts that are expected to be populated with many instances, thus it can be viewed as *information ontology*.
2. The incorporation of class' attributes can be represented as a set of variables and can be stored along with their datatypes. From this point of view the extraction ontology can be used as a data structure, which can come in handy while for example storing the extracted data in database.

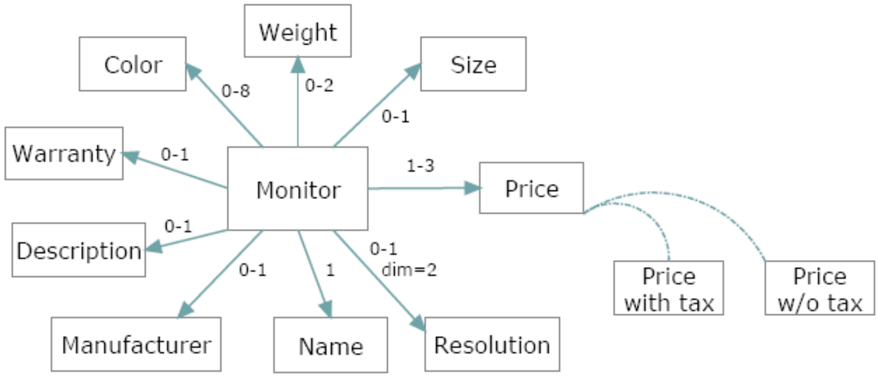


Fig. 1. High-level structure of presentation ontology for computer monitors

3. The extraction ontology can further contain additional higher-level restrictions, such as cardinality or mutual dependency, and therefore it can be looked upon as *knowledge ontology*.

As such ontologies are meant to describe the *presentation* of objects within some media (on the web pages, in the notion of WIE) instead of real-world objects, it is natural to speak about *presentation ontologies*. A presentation ontology represents the fundamental part of an extraction ontology: it is the abstract part that captures the logical structure of the presentation; together with some additional low-level patterns (that enable information extraction) it forms the extraction ontology. In Fig. 1 we see the graphical depiction of a presentation ontology for the computer monitor (product catalogue) domain.

Because of different modeling principles applied while authoring presentation ontologies in contrast to other conceptual models, presentation ontologies have a slightly different nature. Most often they contain a single class, referred to as the *core class*, however multi-class presentation ontologies are also possible. The core class is then supplemented with its attributes and additional constrains. Due to this difference from other knowledge models, a transformation process is needed for their meaningful reuse. Each kind of model has its own specifics; in the remaining sections of this work we will walk through them. However, generally, the transformation process will consist of a few steps that are common regardless of the source of underlying knowledge. These are:

1. choose the core class C and add it to the presentation ontology
2. create its attributes in the presentation ontology
3. formulate ontological constraints (data type, cardinality) over attributes
4. create additional "WIE hooks" for each attribute: in addition to simple datatype restrictions over attributes, more extraction knowledge (e.g. regular patterns) can be added based on the content or context of known or estimated instances.

Regretfully, the structure of domain models can be very variable and the expressiveness of the source models is often high. Such models thus cannot be transformed deterministically, as there are many ways of reusing a single model. Therefore the outcomes of the rules presented below should rather be interpreted as recommendations for an expert designer to help him/her author a suitable presentation ontology.

3 Reuse of Domain Ontologies

While the inclusion of extraction patterns is specific for the WIE setting, the abstract conceptual structure is analogous to that of domain ontologies. As the number of domain ontologies available on the semantic web increases, their reuse would be quite beneficial.

Transformation of a domain ontology expressed in the standard semantic web ontology language OWL[1] (or other high-level ontology language) into a presentation ontology will mainly amount to the transformation steps mentioned above.

The first step is to choose the core class. We so far formulated four rules that can help choose the core class:

a1) Class C that has individuals directly asserted in the domain ontology should probably not become the core class in the presentation ontology.
a2) If some property D does not have an inverse property explicitly declared, a class C in the domain of this property is more likely to become the core class than any class C_1 that figures in its range.
a3) If a class C has a minimum cardinality restriction on property D whose range is class C_1, such that C_1 does not have any restrictions on the inverse property of D, then C_1 should not become the core class.
a4) If there is a chain of object properties $(O_1,O_2, ...,O_n)$, where O_k is object property of C_k, and for every k, $1 \leq k \leq n - 1$ the range of O_k is C_{k+1}, then the classes at the ends of such a chain (i.e. C_1 and C_n) are more likely to form the core class. If a class C is at the end of more such chains, it is even more suitable for becoming the core class.

When a core class is chosen, its attributes have to be created in the presentation ontology. Again these attributes can be based solely on the needs of the presentation ontology creator and on its purpose (in practice they are either chosen ad-hoc or statistically learned from a corpus of sample data), however to maintain the semantic soundness of the resulting data, even the choice of attributes should be based on an existing knowledge model. We thus formulated another set of rules that support the population of the core class C with attributes:

b1) A datatype property may directly yield an attribute. Furthermore a datatype property D of some class C_1, together with a chain of object properties (typically *part-of* properties) $(O_1,O_2, ...,O_n)$, where O_1 is object property of C, O_n is object property of C_1, and for every k, $1 \leq k \leq n - 1$, there is a class having both O_k and

[1] http://www.w3.org/2004/OWL/

O_{k+1} as its properties, may yield an attribute. For example in a weather forecast domain for C = `WeatherForecast`, C_1 = `Weather` and D = `hasTemperature`, and O_1 = `forecastWeather` having C as its domain and C_1 as its range, can yield an attribute such as ‘Temperature’.

b2) A set of mutually disjoint subclasses of C may yield an attribute even without a property counterpart in the source ontology. An example from a weather forecast domain would be C = `Precipitation` having mutually disjoint subclasses `{Rain, Snow, Hailstorm, ...}` can yield the attribute ‘Precipitation’.

b3) A set of mutually disjoint subclasses of some C_1 such that exists a chain of object properties between C and C_1 (in the same sense as in the first rule), may yield an attribute.

b4) An object property O of a class C whose object is a class C_1 that has some individuals asserted in the ontology may yield an attribute. Some possible values of such an attribute could be directly given by the asserted individuals. For example in another ontology from weather forecast domain C = `Weather`, C_1 = `Precipitation` and O = `hasPrecipitation` with C_1 having asserted individual such as `{rain, snow, hailstorm, ...}` can again yield an attribute ‘Precipitation’.

To support these rules we performed tests on freely available domain ontologies with varying domains, size, purpose and structure. Results are shown in Table 1. The table shows how many times each rule could be used while transforming the given ontology into a presentation ontology.

Some of these numbers can be of particular interest, for example it is not hard to spot the correlation between the use of rules b4) and a1), which is due to the fact that both are based on the presence of instances. Further there are three ontologies that were only suitable for rule b3); this can be explained by the fact that they are bare taxonomies. Note that, even if there are no properties present in such taxonomies, it is still possible to derive a few attributes of the core class from them.

It is appropriate to remind once again that these rules are nondeterministic and therefore often identify more core classes, more attributes or none at all. It will always be up to the expert to maintain the notion of the incipient presentation ontology and it will be up to him to refine the results. The numbers in the table only show the rules whose results were really chosen.

Of particular interest are the two large ontologies on which none of the a) rules apply. This is due to the fact that these ontologies cover a broad domain (and are probably meant to be used as a dictionary) and therefore the core class actually chosen by the expert can be distinguished from other classes only with the use of the expert’s innate knowledge. Even if the choice of the core class was not supported by the a) rules, the b) rules can still apply and yield some of its attributes.

In summary, by these tests we verified the possibility of reusing knowledge stored in existing domain ontologies. Next we will focus on other knowledge models that are commonly present in the corporate environment.

Table 1. Transformation rules test results

domain, ontology uri	number of classes	rules a1)	a2)	a3)	a4)	b1)	b2)	b3)	b4)
weather:									
weather-ont (from Semwebcentral)	9	1	4	0	0	4	7	9	1
WeatherConcepts (from LSDIS)	19	0	9	0	6	8	8	3	0
weather-ont3 (from AgentCities)	96	7	38	7	11	0	6	21	6
publication:									
http://www.lehigh.edu/~zhp2/2004/0401/univ-bench.owl	43	0	3	0	3	3	11	4	0
http://sib.deri.ie/fileadmin/documents/swportal.owl	70	0	10	0	4	6	19	3	0
http://www.csd.abdn.ac.uk/~cmckenzi/playpen/rdf/akt_ontology_LITE.owl	61	3	4	4	12	11	25	6	1
http://ebiquity.umbc.edu/ontology/publication.owl	15	0	5	2	0	15	7	1	0
http://alignapi.gforge.inria.fr/tutorial/myOnto.owl	40	1	21	3	4	10	19	10	0
conference:									
http://sib.deri.ie/fileadmin/documents/swportal.owl	70	0	4	0	6	1	22	3	0
http://lsdis.cs.uga.edu/projects/semdis/sweto/testbed_v1_2.owl	43	0	1	0	2	6	9	2	0
http://zeitkunst.org/bibtex/0.1/bibtex.owl	15	0	0	4	0	40	0	1	0
computer:									
http://semweb.mcdonaldbradley.com/OWL/jiva.owl	259	0	2	0	2	2	8	3	0
http://www.openmobilealliance.org/tech/profiles/UAPROF/ccppschema-20021212 (rdfs)	7	0	20	0	0	20	0	0	0
http://morpheus.cs.umbc.edu/aks1/ontosem.owl	7596	0	0	0	0	0	0	5	0
event:									
http://www.ontotext.com/kim/kimo.rdfs	322	0	11	0	4	8	10	4	0
http://rhizomik.net/ontologies/2005/03/TVAnytimeContent.owl	376	0	0	0	0	0	0	32	0
http://smartweb.dfki.de/ontology/swinto0.3.1.rdfs	2006	0	100+	0	44	85	42	14	0
http://www2.sims.berkeley.edu/academics/courses/is202/f04/phone_project/Group8/group78.owl	1650	0	0	0	0	0	0	15	0

4 Reuse of UML Models

Being the standard modeling language in software engineering, UML[2] has received wide attention not only in academic spheres, but also in industrial software development. As a consequence, UML is much better supported in terms of tools and available expertise than the semantic web languages such as OWL. The wide acceptance of UML makes it an ideal candidate for the search for existing knowledge and we think it may prove useful for authoring a presentation ontology. The drawback of trying to reuse knowledge stored in UML models is however their public unavailability, because UML diagrams are often considered the company's precious property. However, for a company such as e-shop wishing to apply information extraction so as to analyze its competitors' offers, it will always be possible to reuse its own models.

According to the Object Management Group specifications the UML diagrams can be divided into several quite different groups: structure diagrams, behavior diagrams and others (such as interaction diagrams). The UML language has been devised in order to integrate competing proposals for modeling languages in the area of software engineering. This integration effort was undertaken in order to push object-oriented design methods into industrial practice. Object oriented design is, in a way, similar to ontological engineering; in some diagrams this is quite obvious. Ontological foundations of UML diagrams are however non-trivial, and worth exploring as e.g. in [5].

There are some ongoing projects that aim to come with a standardized approach of transforming UML diagrams to common ontology languages (for example Falkovych [3]), however they are often only concerned with class diagrams and do not make the best of other parts of UML (which can of course be useful too). Yet the main difference of Falkovych's approach lies in the fact, that in contrast with our work they aim to derive common multi-class domain ontologies, not presentation ontologies.

There are different possibilities of deriving a presentation ontology from diagrams of every group; some guidelines follow.

4.1 Structural Diagrams

The various structural diagrams such as class diagrams, component diagrams, and deployment diagrams describe static, structural constructs (e.g., classes, components or nodes artifacts).

The most common of these diagrams is the *class diagram* because it is very valuable in software engineering tasks. Luckily the concept of class in UML is very similar to the meaning of class in an ontology (and yet it is not the same, for details see [5]).

For our purpose of deriving the presentation ontology we can work with the UML class in the same way as we did with the classes in domain ontologies. Some rules remain intact, but some work differently:

- A class can still directly yield a class in the presentation ontology, and a property can still directly yield an attribute.
- As there are no inverse properties in class diagrams, rule a2) can therefore be seen as even stronger.

[2] http://www.omg.org/technology/documents/formal/uml.htm

- The multiplicity of a relation in a class model can be, with a certain degree of tolerance, translated to a cardinality restriction, and therefore used in rule a3).
- The generalization in class models is nearly equivalent to is-a hierarchy relation, and can therefore serve as support for all chain rules (a4), b1), b3)).
- As multiple inheritance is not commonly allowed in the class model, an individual cannot be instance of more than one bottom-level class and thus all subclasses of a class can be considered as mutually disjoint. This comes in handy in the rules b2) and b3).

We see that the class diagram can be used quite extensively. Similarly, the object diagram fulfils all properties of the class diagram and moreover can incorporate instances. We can use these in a similar way as we did in the case of domain ontologies:

- The instances can be used as individuals in rule a1) for rejecting a core class.
- The instances can be used as individuals in rule b4) for populating the core class with attributes.

Other diagrams in UML that can be considered as structural do not provide such extensive sources for building presentation ontologies, however some of their features may still be useful:

- *Composite structure diagrams* can, along with the information about classes and instances, provide some limited restrictions that can in very specific cases yield an axiomatic rule about existence of some attribute value.
- *Component diagrams* depict the structure of individual components, and therefore the inclusion of a class in some component can be vaguely translated as part-of relation. This relation can again serve as a basis for the chain rules (a4), b1), b3)) or simply yield an attribute as in a1).
- The *package diagram* is used to provide some logical wholes to other diagrams. The existence of a package of some entities that can be mapped to classes can again yield a part-of relation, or sometimes even an attribute. If the package contains some entities that can be mapped to attribute, it could serve some examples of that attribute's values (as a part of the additional extraction knowledge).

However not all kinds of models provide useful information in the means of authoring an extraction ontology. For example the *deployment diagram* contains implementation details, which could prove useful while populating an ontology with individuals, but it is not very helpful for our purpose.

4.2 Behavioral Diagrams

The behavioral diagrams specify the dynamic, behavioral constructs such as activities, interactions, and states. The use of these constructs is not so straightforward as in the case of structural elements, yet they still can be used. One of the most used behavioral diagrams in practice is the *state machine diagram*. The use of the state machine diagram can be supported by the fact that it describes the possible states of every object of a particular class and therefore it can tell something about the class itself:

- The described set of possible states of an entity and transitions between them can yield an attribute in presentation ontology, and moreover it can yield example values of this attribute provided by the individual states.
- Ideally the set of states of an entity described by a state machine diagram is complete, so it should provide a complete enumeration of values of an attribute (provided an attribute is yielded).
- The presence of a choice point in the state machine diagram can be contingent on the existence of a relation to some other entity. This relation can yield an attribute directly or can be used with other rules.

In UML there is also the *activity diagram*, which is basically an extended version of the state machine diagram. The nature of this extension lies in the fact that it can describe dependencies between states of different entities and thus entail some relation/s between them, which can again be further used with the former rules.

Beyond these two diagrams we can place *use case diagrams* in the behavioral group. Use cases are a means for specifying required usages of a system. Typically, they are used to capture the requirements of a system, that is, what a system is supposed to do. The key concepts associated with use cases are actors, use cases, and the subject. The subject is the system under consideration to which the use cases apply, and is often described by the class diagram. The users and any other systems that may interact with the subject are represented as actors. The use case diagrams can also be a source for our approach:

- Actors always model entities that are outside the system and therefore the actor should not yield a core class.
- Although the actor should not yield a core class, it can yield an ordinary class that can further be transformed.
- In the case of actor generalization it is possible to use rules that work with ordinary subclasses.
- In a perfect case the presence of a use case could lead to the existence of a possible state of some entity. Then we could work with the state just like when concerning state machine diagrams.

4.3 Interaction Diagrams and UML Supplements

Interactions are used in a number of different situations. They are used to get a better grip of an interaction situation for an individual designer or for a group that needs to achieve a common understanding of the situation.

Interactions are (according to OMG recommendations) also used during the more detailed design phase, where the precise inter-process communication must be set up according to formal protocols. However, as a source for authoring a presentation ontology the interaction diagrams are very limited, hence the rules for their use are vague:

- Generally the set of interactions between two entities should lead to the existence of at least one relation between them.
- The elements of interaction in individual diagrams can yield possible values of attributes, however these attributes should have been specified elsewhere.

The supplements of UML are only interesting to the extent that they can provide additional information to the extraction part of the final ontology, such as data types of values.

The above is not a complete list of models that exist in UML, however, we see little use of the others at the moment.

5 Suitability of Other Commonly Used Metamodels

UML is not the only framework of metamodels used in today's industry. There are many other various ways of formalizing specific knowledge. Amongst others, two are of particular interest: relational database models and business process models.

The *relational model* for database management is a database model based on predicate logic and set theory, so it also has many things in common with other means of specification of a domain (abundant literature about translating database content into ontology instances exists, e.g. [1]). The reuse of a relational model for building a presentation ontology can also be driven by some non-deterministic rules:

- An entity (i.e. a table) can directly yield a class and its fields (i.e. columns) can directly yield attributes.
- Foreign key references can be used as a general type of concept relation, i.e. a property of a class, and can be used in the chaining rules.
- In contrast to the reuse of domain ontologies it is necessary to distinguish the supporting tables that are incorporating the *m:n* relations. These auxiliary tables should not yield a core class (or any other class), despite they would be rated high by the original (domain-ontology) transformation rules based on property chains.
- Due to the explicit specification of primary and secondary keys it is easy to recognize an inverse property and use the rule a2) if it is not present (and it is not present commonly).
- As inverse properties are not common in the relational model, the chaining rules should be even more effective than in the reuse of domain ontologies.

The *business process model* is designed to describe a collection of activities needed to produce a specific output for a particular customer or market. It is the basic tool of the discipline of business process engineering. A process in this context is a specific ordering of work activities across time and space and it is related to the change of a state of an entity. We can again spot some useful rules:

- Every process depicts a change of a state of some entity and therefore after the fashion of state machine diagrams it should yield a possible value of an attribute, or it can even yield the attribute itself.
- The event element and the choice element should express a relation to some other entity and therefore could yield an attribute given by this relation.

- A set of processes delimited by interactions with external systems (and ideally even enclosed in a pool) should describe an entity, which can then possibly lead to a class in presentation ontology.

6 Related Work

Probably the most similar work to ours is what is presented by Falkovych et. al. in [3]; it is concerned with the transformation of UML class diagrams and relational models to ordinary multi-class domain ontologies. Since with our approach we derive presentation ontologies from various models our works do not intersect very much. The biggest advantage we can take of Falkovych's work is that they manifest and support the similarity of classes and their properties in UML and in domain ontologies.

On the other hand we have to mention the ongoing work on extraction ontologies in general published by Embley in [2] and following, however their aim is exactly opposite to ours: the transformation of extraction ontologies (along with extracted instances) for compliance with existing domain ontologies.

Moreover, our work has a close affinity to projects considering the ontology selection, such as OntoSelect or Watson; because a successful search for ontologies is a necessary prerequisite for our approach, we can extensively use their results.

Part of our work is the Ex project whose main purpose is the development of a WIE tool that can work with concepts such as the ones supplied with an extraction ontology. For further reference, see e.g. [7] and [8].

7 Conclusion and Future Work

We have shown how various knowledge models can be reused in favor of the semantically driven approach to web information extraction. The pros of this reuse are in the fact that the final knowledge model for information extraction is semantically well structured and that the annotation of extracted data is consistent with existing models as much as possible. The cons on the other hand are that the business models are not commonly freely available and so the enterprise has to rely on its own models. Furthermore, there is no standard language for specifying the presentation and extraction ontologies; we use a proprietary format at the moment.

Finally, although it is possible to reuse knowledge stored in any one of the different models, we think it would also be possible to reuse knowledge that is dispersed across multiple models. For example, the role of the state entity in different models could be of particular interest. This is one of interesting directions for further work.

Acknowledgement

The research leading to this paper was supported by the European Commission under contract FP6-027026, Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content - K-Space.

References

1. Astrova, I., Korda, N., Kalja, A.: Storing OWL ontologies in SQL relational databases. In: Proceedings of World Academy of Science, Engineering and Technology (WASET), pp. 167–172 (2007)
2. Embley, D.W., Tao, C., Liddle, S.W.: Automatically extracting ontologically specified data from HTML tables of unknown structure. In: Spaccapietra, S., March, S.T., Kambayashi, Y. (eds.) ER 2002. LNCS, vol. 2503, pp. 322–337. Springer, Heidelberg (2002)
3. Falkovych, K., Sabou, M., Stuckenschmidt, H.: UML for the Semantic Web: Transformation-Based Approaches. In: Knowledge Transformation for the Semantic Web, pp. 92–106. IOS Press, Amsterdam (2003)
4. Gruber, T.R.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition 5(2), 199–220 (1993)
5. Guizzardi, G.: Ontological Foundations for Structural Conceptual Models, Telematica Instituut Fundamental Research Series No. 15 (2005), ISBN 90-75176-81-3
6. van Heijst, G., Schreiber, G., Wielinga, B.: Using Explicit Ontologies in KBS development. Int. J. Human-Computer Studies 46, 183–292 (1997)
7. Labský, M., Svátek, V., Nekvasil, M., Rak, D.: Information extraction using extraction ontologies. In: Proc. PriCKL 2007, ECML/PKDD Workshop on Prior Conceptual Knowledge in Machine Learning and Knowledge Discovery, Warsaw, Poland (2007)
8. Labský, M., Nekvasil, M., Svátek, V.: Towards Web Information Extraction using Extraction Ontologies and (Indirectly) Domain Ontologies. Whistler 18.10.2007 – 21.10.2007. In: K-CAP 2007, pp. s201–s202. ACM, New York (2007)

Community-Driven Ontology Evolution: Gene Ontology Case Study

Anna V. Zhdanova

ftw. Telecommunications Research Center Vienna,
Donau-City-Strasse 1/3.Stock, A-1220 Wien, Austria
zhdanova@ftw.at

Abstract. Communities on the Web capture, represent, and evolve their knowledge using ontologies, either explicitly or implicitly. The Gene Ontology project is a typical and advanced case study of a community-driven ontology creation and evolution. We use this case study to derive and illustrate factors that limit dynamic knowledge sharing in community environments. Specifically, we analyze ontology evolution implemented by the Gene Ontology community over the period of five years, as well as the used infrastructures for knowledge management. We observe limitations of communication practices within community-driven ontology construction, the lack of correlation between requested and actual changes in the ontology, and propose social and technical guidelines for making ontology-based knowledge sharing and evolution more responsive to users' needs.

Keywords: ontology evolution, collaborative semantic environments, gene ontology, community-driven ontology management.

1 Introduction

Community-driven ontology construction has been applied in a number of case studies in various domains, such as in environments to support work of an organization as well as for entertainment and keeping in touch [5, 10, 16] or infrastructures for Semantic Wikis [11, 15]. The Semantic Web approach [2, 6] and development of knowledge portals [8, 13] brought numerous technologies for knowledge representation and sharing that serve as a ground for community-driven ontology evolution. Ontology evolution with social aspects [3] become generally important as there is an emerging need to understand how online communities advance their shared knowledge over time, how to measure and predict the ontology evolution rates, and how these rates co-relate with communities' successes or fruitful collaboration outcomes.

A promising area for community-driven ontology management application comprises environments supporting knowledge-intensive research communities, e.g., in eScience. Typically, life sciences can be seen as an important domain of community-driven Semantic Web application due to large amounts of domain-related information and data that needs to be exchanged between the life scientists. In particular, a charter for "Semantic Web for Health Care and Life Sciences Interest Group" (HCLSIG) has

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 118–129, 2008.

been published by W3C[1]. Community-driven ontology construction is being addressed by the Gene Ontology (GO)[2] developers and users on a large scale. The GO Consortium [4] provides structured, controlled vocabularies and classifications that cover several domains of molecular biology and are freely available for community use in the annotation of genes, gene products, and sequences.

The GO community can be seen as far ahead of other communities in *consensus-grounded* and *collaborative* construction of ontologies [1]. Moreover, the ontology size, its high dynamics rate, years of progressive development, and large number of people involved in the project make the GO community one of the largest, representative and thus important case studies for application of community-driven semantics.

Our research objective in this paper is to observe how the collaborative community-driven ontology evolution is implemented in the real-life practical setting, and thus identify successes and further challenges for the approach. We analyze the way the GO community organizes the ontology construction process, track dynamics of GO over time periods, identify correlations between community involvement and the GO evolution, its up-to-dateness and representation. Using the GO case, we illustrate how the communities' working habits and the state of the art supporting tools can be further advanced towards making the ontology evolution driven primarily by the communities of its users and developers.

The paper is structured as follows. In Section 2, we describe the general relation between goals and usages in Web community infrastructures, and in particular the GO community: their goals and activities. Actual changes taking place in the GO over a five year period are presented in Section 3, and an analysis of these changes is provided in Section 4. In Section 5, we discuss approaches that can assist to evolve the GO. Section 6 concludes the paper.

2 Goals and Activities of the Community

Evidently, in many cases a community is created to reach certain goals, as it is the case for the GO community. At the same time, the reality demonstrates that once the community Web environment starts to run, it is very likely to be used to satisfy also other goals than the ones set by the community environment hosts [12]. Thus "usages" of ontologies in community environments can differ from the creation purpose of these ontologies and environments. In fact, "usages" can redefine the initial community "goals". For example, software developers may find that communities have discovered and are using added-value functionality for a certain purpose whereas the product has not initially been designed for this purpose. The developers can set catering this previously unexpected usage purposes as a goal of further development, thus demonstrating a strong connection between goals and usages.

In Figure 1, we show the main interaction levels we distinguish characterizing goals and usages in community environments, which are as follows.

- Considering *individual user level*: "usage", how people use the community environment,

[1] HCLSIG Charter: http://www.w3.org/2001/sw/hcls/charter.html
[2] The Gene Ontology: http://geneontology.org

- Considering *community level*: explicit and implicit feedback, what people say explicitly and which implicit message they express by interacting with the environment,
- Considering *community maintainers/software level*: “goals”, which goals and purposes the community creators pursue when setting up community infrastructure.

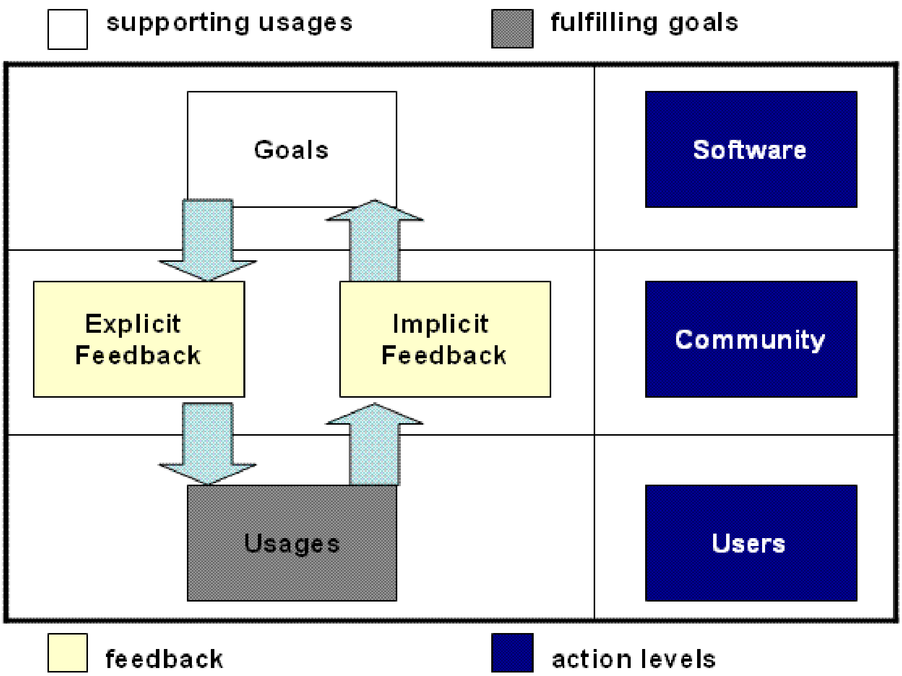


Fig. 1. Goals and usages in community environments

Further, we analyze the GO community adhering to the described above principles. Specifically, the goals, feedback flows and software infrastructure of the GO community are as follows.

Main goals of the GO community are:

- collect, structure, distribute and disseminate information in the field of genomics;
- provide a set of structured vocabularies for specific biological domains that can be used to describe gene products in any organism [4].

The broader goal of Open Biomedical Ontologies (OBO) [14] is to cover the range of biology which is currently described largely in English natural language, and thus facilitate querying, analysis and “de facto” integration.

The GO community reaches its goals and performs its usages employing the following **technical infrastructure**:

- To *collect information*: mailing lists, meetings, SourceForge account;
- To *structure information*: Concurrent Versioning System (CVS), SourceForge account[3], editors such as DAG-Edit[4], formalisms such as OBO language (OBOL);
- To *distribute/disseminate information*: websites geneontology.org, sourceforge.net, CVS, converters to different ontology languages such as to OWL[5].

3 Structure and Changes of GO

Structurally, GO as such consists of three sub-ontologies of the following domains:

- biological process ontology,
- cellular component ontology,
- molecular function ontology.

The molecular function defines what a gene product does at the biochemical level. The biological process normally indicates a transformation process triggered or contributed by a gene product involving multiple molecular functions. The cellular component indicates the cell structure a gene product is part of. As a whole, GO contains around 20 000 concepts.

The ontology structure is fairly simple: GO is a handcrafted ontology accepting only ”is-a” and ”part-of” relationships [14]. The hierarchical organization is represented via a directed acyclic-graph (DAG) structure similar to the representation of Web pages or hypertext systems. Members of the GO consortium group contribute to updates and revisions of the GO. The GO is maintained by editors and scientific curators who notify GO users of ontology changes via email, or at the GO site by monthly reports. When annotating the GO terms, the provided annotations should include its data provenance or source a cross database reference, a literature reference, etc.

GO is available in several different formats, such as: OBO flat file format (obo extension), GO flat file format (ontology extension), XML (RDF/XML) file format (rdf-xml extension), OWL (RDF/XML) file format (owl extension), MySQL Version.

The structure of a GO Term is as follows:

- term name (e.g., “cell”),
- a GO identifier/accession number: an arbitrary (non-semantic, meaningless) unique, zero-padded seven-digit identifier prefixed by GO (e.g., “GO:0005623”),
- optional synonyms (e.g., “synonym of apoptosis= type I programmed cell death”),
- database cross references: identifiers used to maintain cross references among databases (e.g., term “retinal isomerase activity” has the database cross reference EC:5.2.1.3 which is the accession number of this enzyme activity in the Enzyme Commission database),

---

[3] The Gene Ontology Project of SourceForge: http://geneontology.sourceforge.net

[4] DAG-Edit, A controlled vocabulary editor: http://sourceforge.net/project/showfiles.php?group_id=36855

[5] Web Ontology Language: http://www.w3.org/2004/OWL

- definition (e.g., ”The action characteristic of a gene product.”),
- comment (e.g., “Note that this term refers to both the old and new”).

An example of a GO term description in the OBOL 1.2 notation is as follows.

```
[Term]
id: GO:0000015
name: phosphopyruvate hydratase complex
namespace: cellular_component
def: "A multimeric enzyme complex, usually a dimer or an
octamer, that catalyzes the conversion of 2-phospho-D-
glycerate to phosphoenolpyruvate and water." [GOC:jl,
ISBN:0198506732 "Oxford Dictionary of Biochemistry and
Molecular Biology"]
subset: gosubset_prok
synonym: "enolase complex" EXACT []
is_a: GO:0043234 ! protein complex
is_a: GO:0044445 ! cytosolic part
```

All *changes in the gene ontology* are listed explicitly in monthly reports[6], apart from being steadily executed in the shared repositories runtime. The monthly reports contain a concise summary of what has happened in the GO ontologies over the past months, specifically, information about

1. new terms,
2. term name changes,
3. new definitions,
4. term merges,
5. term obsoletions,
6. significant term movements,

as well as general statistical data for the ontologies, such as total quantities of terms for every sub-ontology and the items from the SourceForge tracker that have been closed over the past month.

Addressing the issues of explicit and implicit feedback in the GO community, one has to primarily notice that *active curation* of the GO construction is one of the GO success factors [1]. GO construction is moderated by about 40 GO team members. However, involvement of a broad community of ontology users is limited to their provision of suggestions on ontology modification. Such approach to ontology construction can be seen as restrictive in the light of current consensus modeling solutions which provide community members more opportunities to be involved in ontology construction [16].

Explicit feedback (i.e., what community members request to change) is mainly performed via SourceForge. Specifically, any community member can submit a suggestion on GO modification, e.g., as a “curator request” for issues on the ontology terms. Four categories are offered to choose from when a request is submitted: “new term request”, “other term-related request”, “term obsoletion”, and “none”. Explicit feedback features from sourceforge.org have been available from February 2002 and in March 2002, the first SourceForge requests started to get resolved by the GO curators.

[6] GO community monthly reports: http://www.geneontology.org/MonthlyReports/

In Figure 2, we indicate how many "curator requests" to change the gene ontology were explicitly proposed by the community (the lower line of the graph). As for the ontology evolution as a whole, a steady increase on the work around GO is observed, both in terms added and in relations between these terms. In fact, the number of relations between the terms grows considerably rapidly than the number of terms [9]. On the graph in Figure 2, we summarize the total number of changes in GO (the upper line of the graph) over time. The horizontal axis indicates the time period and the vertical axis indicates the total number of ontology changes.

4 Data Analysis

In this section, we provide the community-driven ontology evolution challenges and limitations derived from the case study. We also discuss the publicly available case study datasets (see Section 2 and 3 for the references) and the specifics of measuring or counting the changes in an ontology developed by a community.

4.1 Challenges and Limitations of Community-Driven Ontology Evolution

Analyzing the GO dynamics data, certain challenges or issues can be identified with respect to the *general community involvement* in ontology construction:

- dynamics of the ontology development only weakly correlates with the development of the actual domain (biology): in particular, at certain points in time substantially more changes are made merely because the curators are more active or because a major formal restructuring of the ontology takes place (as for example, the modification "peak" of April 2003 represented with the upper line of the graph in Figure 2);
- in some cases, actual community involvement goes beyond model organism database communities. For example, the development of the immunology component (comprising 726 new terms) of the GO did not come from any particular model organism database community, but from the immunology community [7]. However, the GO community does not have typical automatized practices for integration of the ontology construction input from adjacent communities. Such integration is performed via ad-hoc modeling and merging, and face to face meetings between the community representatives;
- certain relatively old (e.g., dated from 2002) curator requests are still marked as "open", which shows that the communication process in the community can be improved by employment of an infrastructure allowing support of alternative versions and enabling communities to agree on some parts of these ontologies.

Regarding evaluation of the *quality of community involvement* in the ontology construction and the *completeness and comprehensiveness* of the ontology, the following observations have been derived from the GO community case study:

- implicit feedback (how GO is actually used) is currently directly not considered in ontology construction;

- SourceForge requests from the community are far from directing or causing the majority of changes: as one can see from Figure 2, most of the changes done in the GO are still curator/expert-driven;
- pre-established categories of ontology changes are not equivalently important (e.g., “new terms” are introduced significantly more often than “term merges” take place). Therefore, initial (not user community-driven) categorization of the GO construction operations appears to be an ad-hoc set up. Such pre-categorization and the predefined by experts ontology cannot be comprehensive;
- the effort distribution among the development of ontology items is likely to be inadequate, as long as it is expert, and not community-biased. As the implicit feedback of the user community is not considered, more often demanded ontology items will not necessarily be the ones that are better specified and promoted. Only if new “organisms” formally join the consortium, the specification efforts are triggered. For instance, when plants joined, major ontology evolution work was needed, due to the areas in the domain totally not considered before (even though they have always been relevant): flies, mouse, yeast do not have twigs and leaves, etc.

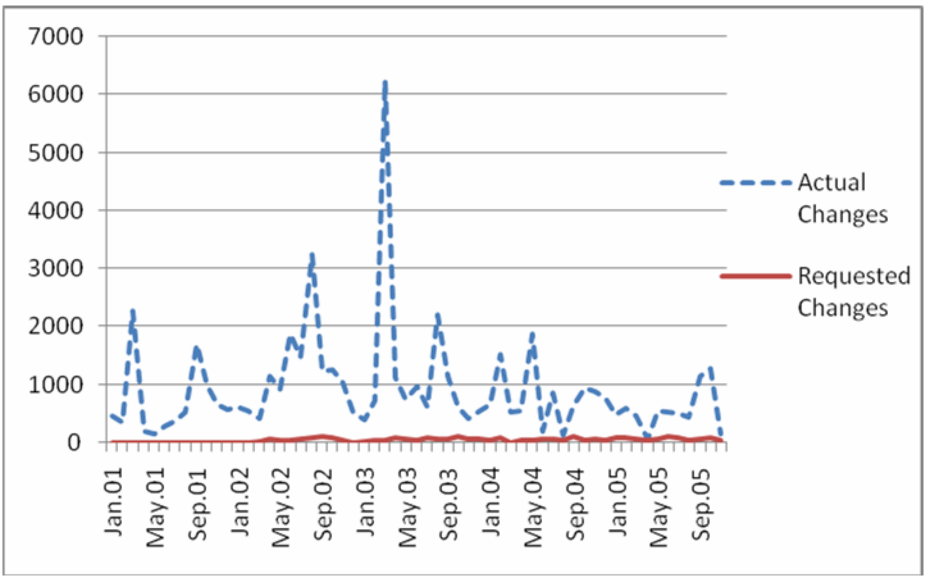


Fig. 2. Total number of changes requested vs. implemented in the GO

4.2 Measuring Community-Driven Ontology Evolution

As indicated earlier, the changes occurring in the GO are classified in six categories: new terms, term name changes, new definitions, term merges, term obsoletions, significant term movements. It should be questioned whether the changes stemming from all the categories are equally important. In particular, one could notice that some changes could be considered as “bulk” modifications that result from a single request and a larger scale operation. For example, the large peak of changes occurring on March 2003 shown in Figure 2 is highly influenced by many homogeneous term name changes in the function ontology: these term name changes constitute ca. 81% of the total changes in the whole GO for this month.

In order to diminish the importance of the "bulk" changes and generally attain a balanced representation of modification rates across different categories, we propose a normalized measure for calculating an "adjusted" weight of total changes in one category over a specific period of time. Using the adjusted weights instead of the weights calculated as the sum of the all the changes provides a semantic (versus "mechanic") view on the ontology evolution. The formula for calculation of the adjusted weight is shown in Equation 1. There, $y_{a,b}$ is a new resulting adjusted weight of the ontology changes made for month a and change category b, and $x_{a,b}$ is an absolute number of such changes (as taken as an input for Figure 2). The total number of time periods (here, months) analyzed in the use case accounts from 1 to n and is denoted as i. The total amount of ontology change categories present in the use case accounts from 1 to m and is denoted as j. The total adjusted weight of changes for a specific month can be obtained as the sum of the adjusted weights of all change categories for this month.

$$y_{a,b} = x_{a,b} \frac{1}{m} \frac{\sum_{i,j=1}^{n,m} x_{i,j}}{\sum_{i=1}^{n} x_{i,b}} \tag{1}$$

Eq. 1. Adjusted weight of ontology changes for month a and category b

The formula is designed to decrease the weights of the change categories where the large number of changes is regularly occurring on a "mechanical" or "bulk" basis and increase the weights of the changes that are performed more rarely and require more community members' inputs and attention. In particular, application of the Equation 1 for month of March 2003 and the category "term name changes", the resulting weight of the changes in this category is 4206 vs. 5052 changes in an un-adjusted weight. For the observed 5 year period, the changes in the category "term name changes" occurred more frequently than average changes in all categories by ca. 20%. The later means that in calculation of the adjusted weights, changes made in this category are counted as less significant than initial number-based weight, while changes in few other categories result in higher adjusted weights than initial weights.

The formula of Equation 1 especially well reflects the situation where the intensiveness of the community feedback needs to be correlated with the intensiveness of the actual ontology development. Different weight adjustment formulas for actual and requested changes can be constructed depending on the needs of the use cases for community driven ontology evolution. Here, as the requested changes the ontology change requests do not normally come in "bulks" but are user-generated on a one-by-one basis, the requested changes numbers are not as substantially influenced by the formula as the adjusted weights for actual changes. However, as the GO use case is not community-driven to its full potential, even after application of the formula and recalculation of the weights of the requested vs. actual changes, the discrepancy between the community-driven and the actual ontology evolution remains.

5 Towards Increased Community Support

As described in Section 3, the explicit feedback to the ontology editing can be partially obtained via ontology development environments like SourceForge. However, current community-driven ontology evolution is hardly influenced by the communities' *implicit feedback*, and we are not aware of tools for integration of such feedback in the ontology evolution process. The contribution of implicit feedback to the ontology construction could be substantially improved by applying emerging social software practices. In particular,

- applications employing the ontology could automatically report the difficulties encountered by the GO users;
- relying on the community in adding, making obsolete or dislocating ontology terms, and not on a curator. The latter involves considerable amount of human labor, long time to perform the change, and the risk of acquiring a single point of failure in the process;
- new and existing ontology items could be automatically suggested to other parties who are potentially interested in these items. For example, as discussed in the community-driven ontology matching approach, the ontology items can be suggested for use to other people under condition that the users belong to relevant communities and social networks [17].

Here we list challenges that are needed to be overcome in the current ontology and knowledge management practices in order to attain the community-driven ontology evolution [16]. In Table 1, we name these challenges in its first column. The requirements on enabling technical infrastructures are listed in the second column of the table, and the requirements on communities' habits are listed in the third column of the table.

Bringing community-driven ontology evolution to the GO and similar communities is targeted at the *following audiences*:

- Developers of various community environments (to illustrate by example the influence of user and developer communities on ontology construction process, and define requirements to the infrastructures allowing benefit from its communities at the highest degree);
- Developers of tools supporting ontology evolution and versioning (to give an idea on which ontology change operations are especially useful and can be successfully captured and processed by the community);
- Computer scientists community, to spot gaps in the market with the case studies for community-driven ontology construction, such as for the GO communities.

In this paper, technical factors of community-driven ontology construction have been considered in more detail than social and organizational factors. This consideration is intentional as (i) the paper mainly addresses the readers with a technical background, (ii) collection and analysis of the social and organizational factors are very complex and go beyond ontology engineering. For instance, a specific influence of the GO community face-to-face meetings or the impacts of the GO advisory board on

Table 1. Habits and infrastructure requirements for community-driven ontology evolution

Challenge	Infrastructure functionality requirements	Community habits requirements
Acquisition of ontologies and annotations	Enable *large scale implicit automatic production of ontologies and semantic annotations* that represent results of users' activities, such as creation and reuse of ontology items in applications, references to existing instance data, etc.	Getting used to *authorized sharing of the data arising from person's activities*; understanding of *security and privacy issues*
Notification and search of ontology items	*Retrieval of relevant ontology items* on the basis of usage histories, personal profiles and preferences, social network and community information	*Learning how to use more complex search, notification and user rules technologies* to their full potential
Ontology visualization and usability	*Visualizing ontology* taking into account importance of the item to a specific user, selection of the *usable presentation mode*	*Understanding of the subjective character of information that is seen*, i.e., why the software demonstrates the ontology and the annotations in the way it does
Versioning support	*Maintenance and usage of the versioning history* in processing of ontology items and annotations	*Getting accustomed to consider the date of the items*, e.g., when searching or designing applications that take into account versioning information
Scalable, distributed infrastructure support	*Accessing and using information from different sources; extracting knowledge from large volumes of information*	Understanding that multiple heterogeneous *information sources can vary in stability, reliability and trustworthiness*
Mobility	*Connectivity between devices* used by community members to work with knowledge; an opportunity to easily switch from one device to another	A habit to *contribute to the community from wherever possible*, in particular, situations not only limited to a typical office setting

the ontology evolution are very difficult to capture, represent and estimate quantity-wise. Once the technical support enables to manage and measure aspects related to social and organizational factors of the community-driven ontology evolution, the impact of such factors would need to be further analyzed.

6 Conclusions

The paper investigates the current community-driven ontology evolution principles, taking the GO community practices as an advanced use case for ontology-based knowledge sharing and evolution. Discrepancies in correlation between the requested by the community vs. the actually performed ontology changes are identified and illustrated. The guidelines for development, measurement and use of advanced infrastructures supporting community-driven ontology evolution are suggested on the basis of the GO case study. As GO is highly community-driven, observations and conclusions drawn from its case shall serve as a flagship for numerous less developed cases of community-driven ontology evolution.

Curator-driven approach made the GO project a success of community-driven ontology development even before the later trend became common with community-driven ontology management, Semantic wikis and Web 2.0 technologies. The restrictiveness of the curator-driven approach keeps the ontology development under control to a higher degree and presumably helps to keep more facts in the ontology scientifically justified or “correct”. However, together with the imperfections of the current ontology development infrastructures the “curated” approach makes the resulting ontology less up-to-date, receiving less user feedback and less complete and representative than it could be. Challenges that are yet to be addressed in the community-driven ontology evolution field are identified, and they generally shall address (i) finding a balance between the “curated” and the purely user-generated content and (ii) combining the technologies that assist to formalize and process more information about the community, including its currently implicit social and organizational factors.

References

[1] Bada, M., Stevens, R., Goble, C., Gil, Y., Ashburner, M., Blake, J., Cherry, M., Harris, M., Lewis, S.: A short study on the success of the Gene Ontology. J. Web Sem. 1(2), 235–240 (2004)

[2] Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American 284(5), 34–43 (2001)

[3] Chen, L., Haase, P., Hotho, A., Ong, E., Mauroux, P.C. (eds.): Proceedings of an ISCW+ASWC 2007 International Workshop on Emergent Semantics and Ontology Evolution (ESOE 2007), Busan, Korea, November 12 (2007)

[4] Creating the gene ontology resource: design and implementation. Genome Research 11(8), 1425–1433 (2001)

[5] Corcho, O., Gomez-Perez, A., Carmen Suarez, M.: The ODESeW platform as a tool for managing EU projects: the KnowledgeWeb case study. In: Staab, S., Svátek, V. (eds.) EKAW 2006. LNCS (LNAI), vol. 4248, Springer, Heidelberg (2006)

[6] Davies, J., Fensel, D., van Harmelen, F. (eds.): Towards the Semantic Web: Ontology-Driven Knowledge Management. John Wiley & Sons, Chichester (2002)

[7] Diehl, A.D., Lee, J.A., Scheuermann, R.H., Blake, J.A.: Ontology development for biological systems: immunology. Bioinformatics 23(7), 913–5 (2007)

[8] Maedche, A., Staab, S., Stojanovic, N., Studer, S., Sure, Y.: SEmantic portAL - The SEAL approach. In: Fensel, D., et al. (eds.) Spinning the Semantic Web, pp. 317–359. MIT Press, Cambridge (2003)

[9] Mungall, C.: Increased complexity of GO (2004), http://www.fruitfly.org/~cjm/obol/doc/go-complexity.html

[10] O’Murchu, I., Breslin, J.G., Decker, S.: Online Social and Business Networking Communities. In: Proc. of ECAI 2004 Workshop on Application of Semantic Web Technologies to Web Communities (2004)

[11] Riehle, D. (ed.): Proceedings of the 2005 International Symposium on Wikis (WikiSym 2005), San Diego, California, USA, October 16-18 (2005)

[12] Shirky, C.: A Group is Its Own Worst Enemy: Social Structure in Social Software. In: Keynote talk at the O'Reilly Emerging Technology Conference, April 24, 2003, Santa Clara, US (2003)

[13] Staab, S., Angele, J., Decker, S., Erdmann, M., Hotho, A., Maedche, A., Schnurr, H.-P., Studer, R., Sure, Y.: Semantic Community Web Portals. Computer Networks 33(1-6), 473–491 (2000)

[14] Stevens, R., Wroe, C., Lord, P., Goble, C.: Ontologies in bioinformatics. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies in Information Systems, pp. 635–657 (2003)

[15] Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., Studer, R.: Semantic Wikipedia. In: Proceedings of the 15th international conference on World Wide Web, WWW 2006, Edinburgh, Scotland, May 23-26 (2006)

[16] Zhdanova, A.V.: Community-driven Ontology Construction in Social Networking Portals. International Journal on Web Intelligence and Agent Systems 6(1) (to appear, 2008)

[17] Zhdanova, A.V., Shvaiko, P.: Community-Driven Ontology Matching. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 34–49. Springer, Heidelberg (2006)

Semantic Integration of Process Models into Knowledge Management: A Social Tagging Approach

Michael Prilla

Ruhr University of Bochum
Information and Technology Management
Universitätsstr. 150
44780 Bochum, Germany
michael.prilla@rub.de

Abstract. Process modelling is an essential task in business and science. Working with process models is a knowledge-intensive task. Unfortunately, tasks like creating, sharing and using process models are hardly supported by knowledge management applications. In this paper, requirements for the integration of process models into knowledge management are described, focusing on the perspective of knowledge work. The paper argues that Social Tagging as a means for semantic content description can fulfil these requirements. Furthermore, an approach of integrating process models into knowledge management and its implementation are shown.

Keywords: Process Modelling, Knowledge Management, Social Tagging, Knowledge Work.

1 Introduction

Process modelling is an essential task in both business and science. There are various applications in which process models are crucial. To name just a few, models are used for analyzing and improving existing processes, designing new procedures in companies and to provide requirements for e.g. developing software.

Work with process models is a complex task. Creating, maintaining and using them not only need an understanding of a model's notation but also demand for domain and contextual knowledge (see [26] for a discussion of this). Thus, tasks concerned with process models have to be regarded as knowledge-intensive. Therefore, process models should be treated as important knowledge artefacts in organizations. In process model related research this has been noticed, leading to various contributions in this field (see section 6 for a short overview). However, there is no comprehensive solution for knowledge management equally handling process models and e.g. textual content. Moreover, further work has to be done in supporting *people* working with process models. Gaining an understanding of these *knowledge workers'* [6], [7] needs is crucial in providing a solution for the problems described above.

In this paper, an approach for a knowledge management solution integrating tasks related to process models as well as equally handling these models in its content

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 130–141, 2008.

management is presented. This approach is based on *bridging the complexity gap* between process models and textual content by *semantic content description*, *integrating* knowledge management related tasks into knowledge workers' tasks and tools and *lowering the usage burden* for these tasks.

The approach presented in this paper contributes in different ways to information system research and business process management. First, it provides a way to lower the usage burden of sharing and using content such as process models. Second, it represents a knowledge management solution equally handling different content types and integrating related tasks into daily work practices. Third, it provides a way in which model usage and the task of modelling itself can be supported. Fourth, from an economic perspective, its focus on supporting knowledge workers tackles one of the most urgent problems in our economy. It should be noted, however, that at the time of writing this paper no evaluation data is available for the approach presented here. Nevertheless, from our point of view it presents a valuable contribution.

In the remainder of this paper, the approach and the corresponding prototypes will be described to a deeper extent. In section 2, the reasons for a knowledge management solution integrating process models will be discussed. After that, requirements for such a solution will be derived in section 3. Section 4 argues that Social Tagging is a mechanism to accomplish this solution. In section 5, the prototype will be described. After that, section 6 discusses related work. The paper concludes with a discussion of the approach and a description of future work on the approach.

2 The Need for Process Models in Knowledge Management

While it is common sense that process models are representations of organizational knowledge, there are no approaches available providing an integrated knowledge management solution being capable of handling process models and other organizational knowledge sources equally. In this section three reasons and constraints are given, which show both the need and potential of such a solution.

2.1 Scarce Usage of Process Models in Organizations

While process models are essential tools in nowadays' organizations, their usage is often limited to a relatively small group of people. This, in turn, results in a situation in which most people in organizations are not aware of certain process models. Therefore, everyday work cannot be compared to standard procedures and existing process models cannot be used for the design or adaptation of new processes. Furthermore, this situation limits the acceptance of process models in an organization.

We came across situations like the ones described above in several field studies and projects. In one of these field studies [28], we worked with companies from German service industry, trying to support them in implementing their processes. When asked for an overview of their processes, they told us that they already had process models describing some of their procedures. However, it turned out that these models were not only out-dated but also only known by a very small group of people in these companies. As a consequence, most processes were implicitly defined for those outside the small group being aware of the process models. Moreover, we encountered situations in which this lead to misunderstandings between process participants, frictions caused by additional coordination efforts and even the re-invention of certain processes due to the unawareness of process definitions [28].

Preliminary results from an ongoing series of interviews started to analyse these observations indicate that the unawareness of process models as information resources leads to a usage barrier towards models and causes their scarce usage.

While measures like teaching people model notations have been taken in these organizations to encourage people to actively use process models, we argue that process model usage depends on the awareness of such models. How are people supposed to use and create models if they do not or even cannot perceive them as valuable information resources? A knowledge management solution integrating common information types such as textual content and information available but scarcely used such as process models may improve this situation.

2.2 Neglecting of Process Models in Knowledge Management

Though knowledge management research has been going on for a long time, existing solutions are still focused towards textual content. There are comprehensive solutions for e.g. mining this type of content. Also, there are widely accepted organizational processes of knowledge management that align tasks related with it towards business processes. Yet, more complex content like process models still plays a subordinate role in knowledge management [26]. From our point of view, this not only worsens the problem described in section 2.1, but also creates additional drawbacks.

First, finding a process model as a knowledge resource for a certain task is hard to accomplish. Considering the empirical observations as described in section 2.1, people either know where to find a certain model or ask other people for it. If the latter task does not lead to success quickly, they go ahead and start work from scratch. This, of course, causes redundancies and bears the risk of incompatible outputs, as there may be multiple versions of the same process not being compatible. Second, the lack of handling process models in knowledge management cuts the possibility of adequately supporting the modelling task with additional information. If process models cannot be related to the contents of a knowledge management application, how can valuable information concerning a model be delivered efficiently to a modeller? Third, even if people happen to find models that may support them in their tasks, without proper knowledge management support they will not be able to find related information providing e.g. details for certain tasks or related processes. Therefore, understanding and using process models in a broader than expert context needs a proper knowledge management solution capable of handling process models.

2.3 Knowledge Work with Process Models

Work with process models is knowledge-intensive. Regardless whether people are concerned with model creation, maintenance or usage, their work has to be regarded as knowledge work [6],[7]. In a setting in which processes change, depend on constraints produced by other procedures and need additional information, supporting these knowledge workers becomes a crucial task. However, applying standard knowledge management solutions for these workers may fail as they often cannot spare the time for activities of knowledge management [6]. Additionally, limiting their freedom in the way of working and in choosing tools to do that work will result in reduced outcomes produced by knowledge workers [6].

Regarding work with process models as knowledge work explains the problems described in section 2.1 partly: The limited time for knowledge management tasks causes artefacts such as process models to be known only to a small group of the creator's network. Spreading awareness of process models therefore depends strongly on her communicational efforts and network. This obviously cuts most people in an organization from the information supply chain. On the other hand, knowledge workers may have personal preferences in information types they use. Therefore, intertwining model-based and textual information also supports them in acquiring knowledge. Recently, the discussion of personal and group oriented information management [2],[8] has produced promising approaches towards efficiently supporting knowledge workers and enabling them to contribute to an organization's knowledge management. The clue for this is the integration of personal information management and organization-wide activities. In other words, with such a solution, knowledge workers contribute to public content by organizing their own content.

3 Requirements

Analyzing the reasons and constraints described in section 2, several requirements can be identified for the integration of content such as process models into knowledge management applications and processes. Besides other requirements mentioned in [26], this paper focuses on those requirements enabling users to participate in such knowledge management tasks and benefit from their participation. Therefore, the overall set of requirements is reduced to four basic demands:

1. **Semantic content description to overcome the complexity gap**
   Textual content and content such as process models can be said to be of different complexity. While the former is linear and coherent, the latter contains ramifications and multiple sub-procedures. Models provide a quicker overview of processes while textual content is more difficult to perceive on first sight. Furthermore, complex content is not as processable to computers as text is. In this paper, this problem is referred to as the *complexity gap* between content types, a phrase borrowed from the so called *semantic gap* in image retrieval research [1],[19]. This gap makes the need of a mechanism to provide homogeneous access to content obvious. Such a mechanism can be provided by semantic content description. Therefore, the foremost requirement for integrating process models into knowledge management is a mechanism for semantic content description.
2. **Low usage burden and high ceiling**
   On the user side, it has always been one goal of information systems research to make the access and contribution of users as easy as possible. However, this must not be done at any price: A low usage burden has always to come with a high ceiling, meaning that the ease to use system must also provide a sufficient surplus in content handling and delivery to users. As can be seen from section 2.3, this consideration especially applies when it comes to supporting knowledge workers. If they are supposed to take part in knowledge management by e.g. sharing process models with other people, this means lowering the usage burden as much as

possible. From a requirements point of view, this also has an impact on the semantic mechanism to be used: Requiring knowledge workers to use advanced and more complex semantic descriptions may not only be incompatible with their mental models [6],[13], but may also distract them from active contribution. Therefore, an easy to use description mechanism should be used.

3. **Integration of all stakeholders**
   The problem of scarce usage of process models in organizations mainly causes two problems. First, knowledge represented by models is not used to the full extent. Second, people knowing organizational procedures best because they perform them every day [31] are not included in the model lifecycle. Therefore, their voice cannot be heard in the management of process models. As a consequence, knowledge management integrating process models must allow for the integration of all stakeholders in an organization and especially for the integration of process participants. This may not only lead to improved knowledge on processes but may also bring together the perspectives of different process participants.
4. **Integration into daily work tasks**
   The habits of knowledge workers imply that their participation in knowledge management strongly depends on the efforts imposed by these tasks. These workers usually stick to their tasks and barely have time for organizing their own information. Therefore, forcing them to use pre-determined processes or tools might result in refusal or at least unmotivated participation in knowledge management. They usually stick to certain tools they find suitable for their work [6]. Therefore, when it comes to sharing content produced by them, this task must be tightly integrated into their daily work practices. Thus, support for e.g. describing and centrally storing content should be built into these tools. Remembering Grudin's still true notion of problems in Groupware, *"the disparity between those who to the work and those who get the benefit"* [14], this integration should also be beneficial for the tasks performed in these tools. For process modelling, this means that model description and sharing should be supported in modelling application. Additionally, the task of sharing models should allow for direct feedback showing people the beneficial effect of sharing their content.

4 Approach: Social Tagging and Tool Integration

Considering the notion of bridging the complexity gap between content types, the question remains how semantic descriptions for this can be realized. There are many semantic technologies available, ranging from ontologies [12] to metadata-based approaches such as Social Tagging [11],[25]. While all of these techniques provide benefits in several applications, for the approach Social Tagging was chosen.

Briefly, Social Tagging consists of adding free-form keywords – "Tags" – to resources without any constraint. Though this may be perceived as an unstructured method with minor quality outcomes at first sight, studies in Social Tagging show that it is well suited as a semantic content description technique [5],[11],[24]. This can be especially seen by looking at the emergence of semantic structures in Social Tagging. In several studies, it has been shown that tags converge to a socially shared and meaningful vocabulary [11],[20] which represents conceptual hierarchies and is not affected by noise from different tags [29]. However, the unrestricted use of tags is often criticized to bear risks caused by homonymy, polysemy and synonymy [1],[11]: similar tags may have different meanings, one tag may be used in different contexts and therefore have different meanings in these contexts and different tags are used to denote the same meaning. Recent contributions show that if these problems occur they can be addressed by clustering algorithms [4] or combining tags with other structural characteristics [1] in order to disambiguate the vocabulary.

The suitability of Social Tagging for the organization of process models is described in [26]. Its general applicability for tasks of knowledge management has been discussed and shown by e.g. [22],[29].

While these arguments identify Social Tagging as a viable candidate for integrating process models into knowledge management content and activities, for a choice of a semantic description mechanism a closer look at the requirements described in section 3 is necessary (omitting requirement 4, as it will be covered in section 5). For the sake of brevity, the discussion will focus on comparing ontologies as the state-of-the-art semantic technique in knowledge management and Social Tagging[1].

Regarding the first requirement of *semantic descriptions* to overcome the complexity gap induced by content types in knowledge management, no decision between ontologies and Social Tagging can be made. Both approaches have shown that they produce valuable content descriptions in practice (e.g. [12],[29]). Moreover, both approaches are capable of handling different content types [16],[33]. This changes when it comes to the second requirement of providing a *low usage burden*. Ontologies are complex constructs built foremost for machine understanding of descriptions. Therefore, they contain hierarchies and rules determining their elements' meaning [12]. This, however, imposes an extra learning effort on people using ontologies [13],[25]. In contrast, the application of free-form keywords to resources can be done without any learning effort. Regarding the notion of a high ceiling, both approaches have been shown to provide valuable descriptions of content. But, concerning the notion of a low usage burden, Social Tagging fits better. For the third requirement of *integrating relevant stakeholders* into both describing and benefitting from process model enabled knowledge management, Social Tagging should be preferred over ontologies, too. This is because ontology descriptions of content are usually done by experts due to the complexity of ontological hierarchies and rules [13]. In contrast, Social Tagging provides a bottom-up approach giving all participants equal voices. It therefore is better suited for integrating stakeholders.

Summing up, for the basic requirements related to content description described in section 2, Social Tagging seems to fit better than other mechanisms such as ontologies. It provides sufficient quality in content description, is easy and quick to use and also integrates relevant stakeholders with equal weight. Therefore, Social Tagging was chosen for the equal handling of process models and textual content in knowledge management.

5 A Prototypical Infrastructure

To show the capability of the approach to integrate process models into knowledge management and foster creation, sharing, management and usage of models a

---

[1] It should be noted that the following arguments also apply to most alterative solutions as these show characteristics lying between those of ontologies and Social Tagging.

prototypical infrastructure was implemented. Recurring to the requirement 4 from section 3, two measures were taken. First, the infrastructure uses already existing tools in order to let people work with tools they are used to. To fully integrate the approach into daily work tasks of modellers or people seeking information, our knowledge management application Kolumbus 2 [27] and our process modelling tool for the SeeMe modelling notation [17] were extended by tagging mechanisms. Second, to enable content sharing, these applications are connected via web services. This exchange is done via an XML representation of SeeMe models, including their tags (cf. [26]). In what follows, a brief walkthrough covering the functionality of the prototype is described with respect to the requirements described in section 4.

The tagging functionality of the modelling editor is described in [26]. Therefore, the walkthrough starts with a tagged process model as shown in Fig. 1.

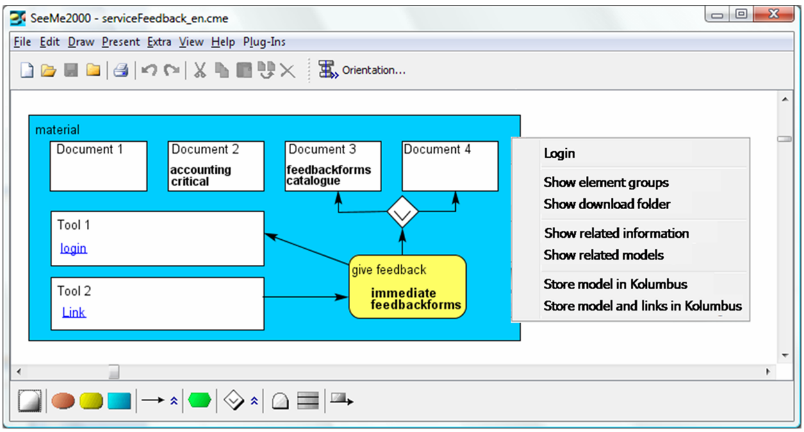


Fig. 1. Sharing process models in a modeling editor

In the prototype, sharing the model is integrated into the modelling application. As can be seen in Fig. 1, a modeller can log into a knowledge management application – if no automatic login is chosen – by a context menu entry. Besides other functions, she can then choose to share only her model and its tags (`Share model in Kolumbus`) or to share her model and all information linked to it, including e.g. URLs (`Share model and links in Kolumbus`). After choosing one of these functions, the content is shared in Kolumbus 2, as can be seen in Fig. 2.

As shown in Fig. 2, not only the model (`serviceFeedback_en.cme`), but also its associated links (`login Tool 1, Link Tool 2`) as shown in Fig. 1 are shared in the knowledge management application. Additionally, the tags applied to the model and its elements are processed by Kolumbus 2 and used as content descriptors. This can be seen in Fig. 3, which shows an alternative view on the content space with several information units tagged similarly to the process model from Fig. 1.

Fig. 2. Shared model and linked content in Kolumbus 2

In Kolumbus 2, content from different areas of the system can be found by using a tag as a search string or using a so called tagcloud. The result of such a query is represented in a tag-based content view, which is shown in Fig. 3. From figures 2 and 3, it can be seen that content is contextualized in two ways: It is represented in a content tree as well as linked to other content via tags. Queries can also be performed from the SeeMe modeling editor by using e.g. the function `Show related information` shown in Fig. 1. Assuming a modeler completing a model and needing additional information on certain elements, the function call in the modeling application will provide her with a list of similar content available in the knowledge management application. Note that similarity here is based on similar tags. An example of a list showing similar information is shown in Fig. 4.

This quick walkthrough presenting the prototypical implementation of our approach demonstrates that is is capable of realizing the contributions mentioned in section 1. First, it enables users to centrally store and therefore share process models with a low usage burden. Furthermore, this task is integrated into tools users are used

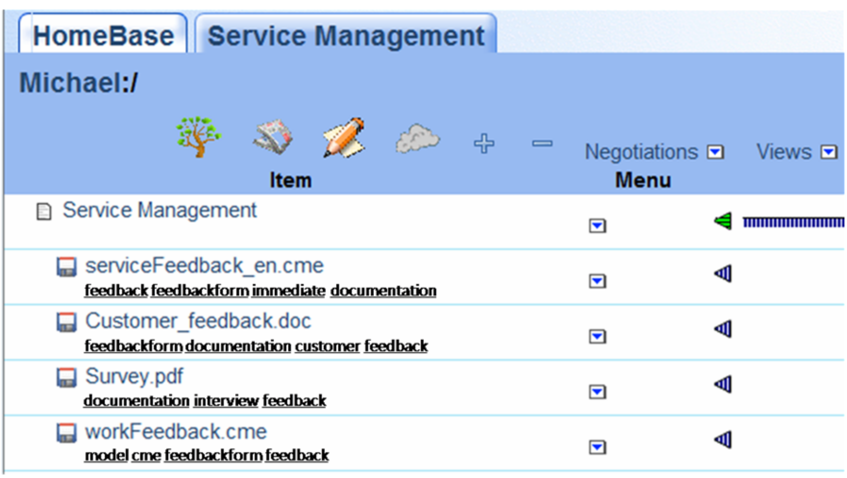


Fig. 3. Tagged content in Kolumbus 2

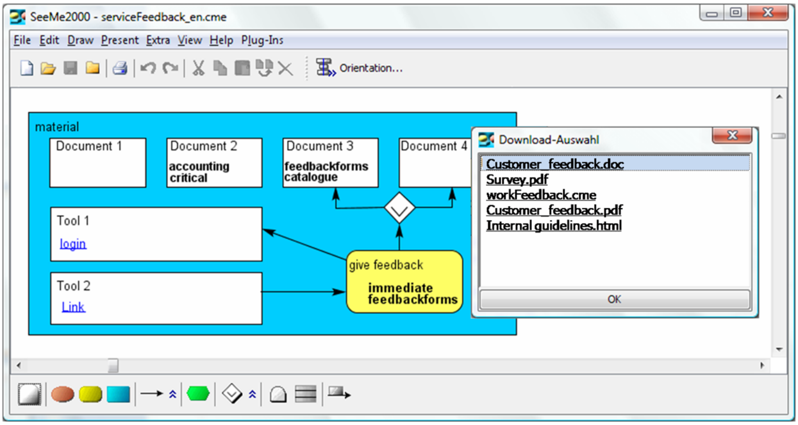


Fig. 4. Related information in the modeling editor

to. Second, the walkthrough shows that model-based and other content is handled equally in the approach as can be seen from e.g. Fig. 3. Third, the setting represented by Fig. 4. shows that the approach also supports the modeling task itself by providing helpful content. Forth, the combination of tool integration and a low usage burden fulfils the needs imposed by knowledge workers.

6 Related Work

There are several research findings related to the approach presented in this paper. In the following, some areas of influence are briefly sketched, focusing on efforts in managing process models. After that, the differences of our approach and its potential combination with existing approaches are described.

Process and model repositories: Reusing models provides a means to use existing processes in newly combined processes. Among these approaches, "online libraries" to support this task [21] and criteria for building up process model catalogues [10] are proposed. Another example can be found in the Google 3D Warehouse[2].

Process-oriented knowledge management: Process-oriented knowledge management uses processes to structure the content of knowledge management systems. This way, knowledge can be provided for specific tasks [23] and process models can be used to support browsing the content [18].

(Semantic) Business process management: Business Process Management [30] mainly focuses on the management of process execution and monitoring. Semantic approaches applying ontologies as described in e.g. [16] aim at enhancing existing systems.

Process model management and maintenance: Approaches supporting the management of process models aim at collaborative creation of process models and their exchange. Examples of approaches can be found in [3] and [32].

[2] http://sketchup.google.com/3dwarehouse

Integration of Heterogeneous Content in Knowledge Management: There are several approaches addressing the problem of integrating heterogeneous content into knowledge management. Most of these approaches apply ontologies as semantic descriptions for such content [9]. Existing approaches address e.g. the integrating of audiovisual content [15].

The approaches presented above provide appropriate solutions to the problems they address. However, they do not provide an integration of process models into knowledge management and the support for tasks related to knowledge work. Moreover, most approaches use formal description schemes such as ontologies, which have been identified as less applicable for knowledge workers compared to Social Tagging. Besides the advantages discussed in section 4, there is an additional advantage in the approach presented in this paper: Our approach focuses on an overall solution of integrating textual and model-based content into knowledge management activities, which enables users to find information in a single system instead of separate systems for organizing processes and textual content. Thus, by interrelating textual and model-based solutions the choice is on the knowledge worker to use the type of information codification suiting her best. Furthermore, our approach connects tools users are used to with knowledge management applications. Therefore, sharing and using information takes place in and from these tools and therefore frictions and barriers caused by switching applications are minimized.

However, we do not expect our approach to solve all problems occurring in the management of processes and are aware that related approaches tackle slightly different problems. Therefore, there is strong synergy potential in combining the approach with e.g. business process management or process oriented knowledge management.

7 Discussion and Further Work

In this paper, an approach in integrating process models into knowledge management is presented. The underlying concept of our approach is based on the requirements described in section 2. The requirements for this approach are derived from empirical and theoretical observations in process model usage, process model handling in knowledge management, knowledge workers and knowledge acquisition. Based on these requirements, it is argued that semantic content description with Social Tagging can be a means to bridge the complexity gap between model-based and textual information. Furthermore, a prototype implementing the concept and is described and the benefits of its usage in terms of supporting creation, sharing, usage and maintenance of process models is shown.

From an economic point of view, neglecting process models in knowledge management means wasting resources and time to produce them. Though it is hard to measure the benefit created by our approach in monetary or time-related metrics, it can be stated that it can foster the usage of now scarcely used model-based information resources. Furthermore, the focus on supporting knowledge workers' needs tackles a current problem. Regarding knowledge workers as an important pillar of our current economy, supporting their needs and enabling them to share and use more information sources can be seen as economically meaningful and beneficiary.

The approach presented in this paper is work in process. Although it is based on a theoretical and empirical basis, at the time of writing this paper we did not empirically scrutinize or measure its benefits yet. Though at the time of writing this paper therefore no evaluation data is available, we are convinced that it is capable of tackling the problems described in this paper and establish process models as more often used knowledge resources.

In further work on this approach, the focus will be set on getting empirical data on the effect the approach has on knowledge work with process models as well as on knowledge acquisition. This will be done in upcoming field studies and experiments. Based on the results of this evaluation, the prototype will be enhanced by e.g. tag clustering mechanisms supporting tag-based content description. Furthermore, the number of applications capable in participating in the overall solution will be enlarged.

References

1. Aurnhammer, M., Hanappe, P., Steels, L.: Integrating Collaborative Tagging and Emergent Semantics for Image Retrieval. In: Collaborative Web Tagging Workshop at WWW 2006, Edinburgh (2006)
2. Barreau, D.: Context as a factor in personal information management systems. Journal of the American Society for Information Science 46, 327–339 (1995)
3. vom Brocke, J., Thomas, O.: Designing Infrastructures for Reusing Conceptional Models - A General Framework and its Application for Collaborative Reference Modelling. In: Abramowicz, W., Mayr, H.C. (eds.) Business Information Systems. 9th International Conference on Business Information Systems (BIS 2006), Köllen (2006)
4. Brooks, C., Montanez, N.: Improved annotation of the blogosphere via autotagging and hierarchical clustering. In: Proceedings of the 15th international conference on World Wide Web, pp. 625–632 (2006)
5. Cattuto, C., Loreto, V., Pietronero, L.: Collaborative Tagging and Semiotic Dynamics. In: PNAS, vol. 104, pp. 1461–1464 (2007)
6. Davenport, T.: Thinking for a living. Harvard Business School Press (2005)
7. Drucker, P.: Landmarks of Tomorrow: A Report on the New Post-Modern World (1959)
8. Erickson, T.: From PIM to GIM: personal information management in group contexts. Communications of the ACM 49, 74–75 (2006)
9. Fensel, D., van Harmelen, F., Klein, M., Akkermans, H., Broekstra, J., Fluit, C., van der Meer, J., Schnurr, H., Studer, R., Hughes, J.: On-To-Knowledge: Ontology-based Tools for Knowledge Management. In: Proceedings of the eBusiness and eWork, pp. 18–20 (2000)
10. Fettke, P., Loos, P., Zwicker, J.: Business Process Reference Models: Survey and Classification. In: Proceedings of the Workshop on Business Process Reference Models (2005)
11. Golder, S., Huberman, B.: The structure of collaborative tagging systems. Journal of Information Science, 198–208 (2006)
12. Gomez-Perez, A., Fernandez-Lopez, M.: Ontological Engineering. With Examples from the Areas of Knowledge Management, E-Commerce and the Semantic Web. Springer, Heidelberg (2004)
13. Grudin, J.: Enterprise Knowledge Management and Emerging Technologies. In: Proceedings of the 39th Annual Hawaii International Conference on System Sciences. HICSS 2006, vol. 3 (2006)
14. Grudin, J.: Why CSCW Applications fail: Problems in the Design and Evaluation of Organizational Interfaces, pp. 85–93. ACM, New York (1988)

15. Hammiche, S., Benbernou, S., Hacid, M., Vakali, A.: Semantic retrieval of multimedia data. In: 2nd ACM international workshop on Multimedia databases, pp. 36–44 (2004)
16. Hepp, M., Roman, D.: An Ontology Framework for Semantic Business Process Management. In: Proceedings of Wirtschaftsinformatik 2007 (2007)
17. Herrmann, T., Hoffmann, M., Kunau, G., Loser, K.: A modelling method for the development of groupware applications as socio-technical systems. J. Behaviour and Information Technology 23, 119–135 (2004)
18. Hoffmann, M., Diefenbruch, M., Goesmann, T., Herrmann, T.: PRomisE2 - Recording and Displaying Situated Process In-formation in Knowledge Management Applications. In: Proceedings of I-KNOW 2002 (2002)
19. Jorgensen, C.: Image Access, the Semantic Gap, and Social Tagging as a Paradigm Shift. In: 18th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research, Milwaukee, Wisconsin (2007)
20. Kipp, M., Campbell, D.G.: Patterns and Inconsistencies in Collaborative Tagging Systems: An Examination of Tagging Practices. In: Proceedings Annual General Meeting of the American Society for Information Science and Technology (2006)
21. Lang, K., Taumann, W., Bodendorf, F.: Business Process Reengineering with Reusable Reference Process Building Blocks. In: Scholz-Reiter, B., Stickel, E. (eds.) Business Process Modelling (1996)
22. MacGregor, G., McCulloch, E.: Collaborative tagging as a knowledge organisation and resource discovery tool. Library Review 55, 291–300 (2006)
23. Maier, R., Remus, U.: Towards a framework for knowledge management strategies: process orientation as strategic starting point. In: Proceedings of the 34th Annual Hawaii International Conference on System Sciences (2001)
24. Marlow, C., Naaman, M., Boyd, D., Davis, M.: HT 2006, tagging paper, taxonomy, Flickr, academic article, to read. In: Proceedings of the seventeenth conference on Hypertext and hypermedia, pp. 31–40 (2006)
25. Mathes, A.: Folksonomies - Cooperative Classification and Communication Through Shared Media. In: Doctoral Seminar of Computer Mediated Communication (2004)
26. Prilla, M., Herrmann, T.: Semantically Integrating Heterogeneous Content: Applying Social Tagging as a Knowledge Management Tool for Process Model Development and Usage. In: Proceedings of I-KNOW 2007 (2007)
27. Prilla, M., Ritterskamp, C.: Collaboration Support by Co-Ownership of Documents. In: Hassanaly, P., Herrmann, T., Kunau, G., Zacklad, M. (eds.) Proceedings of COOP 2006, IOS Press, Amsterdam (2006)
28. Schermann, M., Prilla, M., Krcmar, H., Herrmann, T.: Bringing life into references process models: A participatory approach for identifying, discussing, and resolving model adaptations. In: Proceedings of Mulitkonferenz Wirtschaftsinformatik (MKWI 2008) (2008)
29. Schmitz, C., Hotho, A., Jäschke, R., Stumme, G.: Kollaboratives Wissensmanagement. In: Pellegrini, T., Blumauer, A. (eds.) Semantic Web. Wege zur vernetzten Wissensgesellschaft (2006)
30. Smith, H., Fingar, P.: Business Process Management: The Third Wave. Meghan-Kiffer Press (2003)
31. Suchman, L.: Office Procedure as Practical Action: Models of Work and System Design. ACM Transactions on Office Information Systems 1, 320–328 (1983)
32. Thomas, O., Scheer, A.: Tool Support for the Collaborative Design of Reference Models? A Business Engineering Perspective. In: System Sciences, 2006. HICSS 2006. Proceedings of the 39th Annual Hawaii International Conference on System Sciences (2006)
33. Zollers, A.: Emerging Motivations for Tagging: Expression, Performance, and Activism. In: Tagging and Metadata for Social Information Organization Workshop, WWW 2007 (2007)

Influence Factors of Understanding Business Process Models

Jan Mendling[1] and Mark Strembeck[2]

[1] BPM Cluster, Faculty of Information Technology
Queensland University of Technology, Australia
j.mendling@qut.edu.au
[2] New Media Lab, Institute of Information Systems
Vienna University of Economics and Business Administration, Austria
mark.strembeck@wu-wien.ac.at

Abstract. The increasing utilization of business process models both in business analysis and information systems development raises several issues regarding quality measures. In this context, this paper discusses understandability as a particular quality aspect and its connection with personal, model, and content related factors. We use an online survey to explore the ability of the model reader to draw correct conclusions from a set of process models. For the first group of the participants we used models with abstract activity labels (e.g. A, B, C) while the second group received the same models with illustrative labels such as "check credit limit". The results suggest that all three categories indeed have an impact on the understandability.

1 Introduction

Even though workflow and process modeling have been used extensively over the past 30 years, we know surprisingly little about the act of modeling and which factors contribute to a "good" process model in terms of human understandability. This observation contrasts with the large body of knowledge that is available for the formal analysis and verification of desirable properties, in particular for Petri nets. To guarantee a certain degree of design quality of the model artifact in a wider sense, several authors propose guidelines for the act of modeling [1,2] but yet with little impact on modeling practice. Clearly, an empirical research agenda is required for acquiring new insights into quality [3] and usage aspects [4] of process modeling.

Following this line of argument, a recent empirical study provides evidence that larger, real-world process models tend to have more formal flaws (such as deadlocks) than smaller models [5,6]. One obvious hypothesis related to this phenomenon would be that human modelers lose track of the interrelations of large and complex models due to their limited cognitive capabilities [7], and then introduce errors that they would not insert in a small model. There are other factors beyond size that presumably affect the understandability of a process model such as the degrees of sequentiality, concurrency, or structuredness [8].

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 142–153, 2008.

Validating such hypothetical relationships empirically would not only represent a major step forward towards understanding quality of process models beyond verification, but also provide a sound theoretical basis for defining guidelines for process modeling in general.

Since only little research has been conducted on quality aspects of process models so far [3,9], we approach this area with an experimental design focusing on the understandability of process models (*not* of process modeling languages). By having a online questionnaire filled out, we aim to gain insight into empirical connections between personal and model characteristics and the ability of a person to understand a process model properly. In particular, we want to find out how the textual content of the activity labels might influence understandability. Figures 1 and 2 show two process model variants that were included in the questionnaire.

Our contribution related to process model understandability is twofold. First, we operationalize understandability and identify three categories of factors, i.e. personal, structural, and textual, that potentially influence model understandability. Second, we present the findings from an experiment that tested the relative importance of these factors. Indeed, all three categories appear to be relevant according to a logistic regression model for the experiment data. Against this background, the remainder of the paper is structured as follows. In Section 2 we discuss related work and identify a lack of empirically validated insight on the understandability of process models. Then, Section 3 introduces the research design, in particular, the conceptualization of the questionnaire and the data that we gather. In Section 4 we present the results of the statistical analysis. Section 5 concludes the paper, discusses limitations of the findings, and identifies open questions that need to be addressed by future research.

2 Related Work on Understandability

There are basically three streams of research related to our work in the conceptual modeling area: top-down quality frameworks, bottom-up metrics related to quality aspects, and empirical surveys related to modeling techniques.

One prominent *top-down quality framework* is the SEQUAL framework [10,11]. It builds on semiotic theory and defines several quality aspects based on relationships between a model, a body of knowledge, a domain, a modeling language, and the activities of learning, taking action, and modeling. In essence, syntactic quality relates to model and modeling language; semantic quality to model, domain, and knowledge; and pragmatic quality relates to model and modeling and its ability to enable learning and action. Although the framework does not provide an operational definition of how to determine the various degrees of quality, it has been found useful for business process modeling in experiments [12].

The Guidelines of Modeling (GoM) [2] define an alternative quality framework that is inspired by general accounting principles. The guidelines include the six principles of correctness, clarity, relevance, comparability, economic efficiency, and systematic design. This framework was operationalized for Event-driven

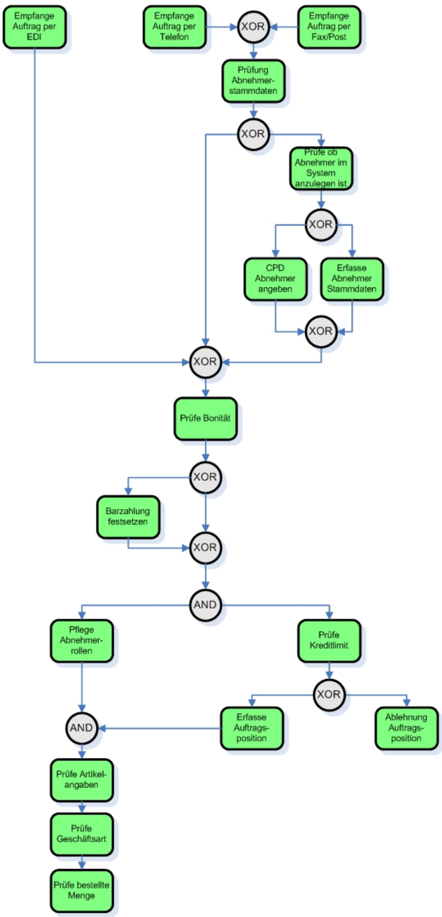


Fig. 1. Model 4 with Text (in German)

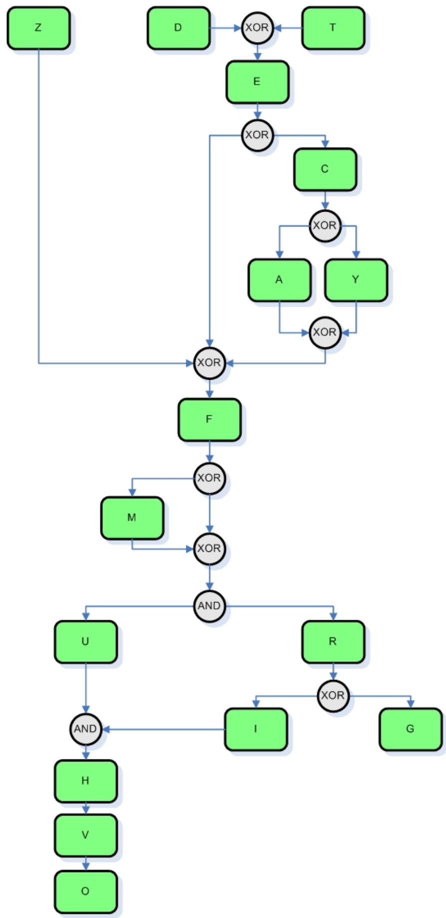


Fig. 2. Model 4 with Letters

Process Chains (EPCs) and also tested in experiments [2]. Furthermore, there are authors [3] advocating a specification of a quality framework for conceptual modeling in compliance with the ISO 9126 standard for software quality [13]. A respective adaptation to business process modeling is reported in [14]. Our experiment addresses partial aspects of these frameworks. In particular, we focus on understandability of process models as an enabler of pragmatic quality (SEQUAL) and clarity (GoM). This requires us not only to ask about understandability, but also check whether models are interpreted correctly. This is in line with research of [15] who experimented on conclusions that people can draw from models. In a more general design setting, the work on cognitive dimensions stresses understanding as one important interaction of a user with a model [16].

Several authors have published work on *bottom-up metrics related to quality aspects* of process models, stemming from different research and partially isolated from each other [17,18,19,20,21,22,23,24,25] , or see [8] for an overview. Several of these contributions are theoretic without empirical validation. Most authors doing experiments focus on the relationship between metrics and quality aspects: [23] study the connection mainly between count metrics – for example, the number of tasks or splits – and maintainability of software process models; [26] validates the correlation between control flow complexity and perceived complexity; and [6,8] use metrics to predict control flow errors such as deadlocks in process models. The results reveal that an increase in size of a model appears to have a negative impact on quality. This finding has an impact on the design of our questionnaire. To gain insights that are independent of process size, we keep the number of tasks constant and study which other factors might have an impact on understandability.

Finally, there are some *empirical surveys* related to modeling techniques. [27] study how business process modeling languages have matured over time. While this is valuable research it does not reveal insights on single, concrete process models. The same holds for [28] who study the usability of UML. [29] approach understandability, not of individual process models, but on the level of the modeling language. They find that EPCs seem to be more understandable than Petri nets. Inspired by this survey we decided to use an EPC-like notation in our questionnaire to minimize the impact of the notation on understandability.

To summarize, there is essentially one relation that seems to be confirmed by related research, and that is that larger models tend to be negatively connected with quality. The aim of our questionnaire is to enhance this rather limited body of knowledge.

3 Research Design

Related to understandability, we identify the following six research questions related to the factors that might influence understandability of process models [29,8,30,11,9]:

1. What *personal* factors (beyond general psychological and intellectual factors) have an influence?
2. Which *model* characteristics (e.g. number and type of splits) contribute to a good understandability?
3. How is understandability related to the *textual content* that is described in the model?
4. How does the modeling *purpose* (e.g. documentation versus enactment) relate to understandability?
5. Which differences in understandability exist when observing semantically equivalent models described in different *modeling languages*?
6. What is the impact of different *visual layout* strategies or graph drawing algorithms on understandability?

We approach these questions with an experimental design focusing on personal, model, and content characteristics (question 1, 2, and 3). Furthermore, we strive to neutralize the influence of the other factors: related to question 4, we gathered a set of process models from practice that capture different domains such as order processing and price calculation. All models were created for documentation purposes (question 4). Based on the observation by [29] that EPCs appear to be easier to understand than Petri nets, we chose an EPC-like notation without events (question 5). The participants received a short informal description of the semantics similar to [31, p.25]. Finally, we drew all models in the same top-to-bottom style with the start element at the top and end element at the bottom (question 6).

The experiment was conducted in three phases. First, we collected a set of six process models from practice that could be displayed on an A4 page. For each of these models we constructed a variant where the activity labels were replaced by capital letters as identifiers. The models were similar to model 4 depicted in Figures 1 and 2. For the 6 models we identified 6 yes/no questions related to the structure and the behavior specified by the model. These questions together with questions on personal experience and knowledge of process modeling were packed into two variants of the questionnaire, one for models with textual activity labels, one for models with letters. Second, we developed a website for conducting the survey as an online questionnaire. We chose an online questionnaire to get practitioners with modeling experience involved more easily. Furthermore, we were able to record the answer times, randomly define the presentation order of the 6 models, and we could randomly assign a questionnaire variant to the participant. Participation was voluntary. As an incentive the participants received feedback about their performance. Finally, the data was analyzed using the statistical software package SPSS.

3.1 Data Gathered in the Survey

In the survey we gathered the following data related to the participants:

- THEORY: The participants had to answer six theoretical yes/no questions without before seeing the models about selected topics related to process modeling such as choices, concurrency, loops, and deadlocks. THEORY captures the sum of correct answers to these questions.
- DURATION: The participants were asked for how long they have been involved with business process modeling. The variable was measured ordinally on four levels: less than one month, less than a year, less than three years, and longer than three years.
- INTENSITY: The participants had to indicate how often they work with process models. There were four options to answer: daily, monthly, less frequent than monthly, never.
- TIME: This variable measures the time that the participants invested in answering the questionnaire.
- TEXT: This variable indicates whether the activities had textual labels (value 1) or only abstract letters (value 0).
- PSCORE: This variable is calculated based on the answers given by the participant to the model related questions. It captures the number of correct

answers by the person. The maximum value is 36 for six questions on six models. This variable serves as an operationalization of understandability related to a person.

Furthermore, we calculated some model metrics from the set proposed by [8]. These include:

- SIZE: This variable refers to the number of nodes of the process model graph.
- DIAMETER gives the length of the longest path from a start node to an end node in the process model.
- STRUCTUREDNESS of the process graph is one minus the number of nodes in structured blocks divided by the number of nodes.
- SEPARABILITY relates the number of cut-vertices to the number of nodes.
- TOKEN SPLIT sums up all concurrent threads that can be activated by AND-splits and OR-splits in the process.
- CYCLICITY relates number of nodes on cycles to all nodes.
- HETEROGENEITY gives the type entropy of the connectors.
- SOUND indicates whether the process model is sound according to [32].
- MSCORE: This variable is calculated based on the answers given by the participants to the model related questions. It captures the sum of correct answers for this model. This variable serves as an operationalization of understandability related to a model.

Finally, we also measured aspects related to the textual labels of the model and correct answers to individual questions:

- TEXTLENGTH gives the string length of all textual activity labels in the process model.
- CORRECTANSWER captures for each individual question answered by a participant whether it was answered correctly (value 1) or not (value 0). This variable serves as an operationalization of understandability related to a model aspect.

We distributed the link to the experiment via the German mailing lists EMISA and WI as well as among students that followed courses on process modeling at the Vienna University of Economics and Business Administration. Typically both academics and practitioners with an interest in conceptual modeling and information systems development are registered with these lists. The questionnaire was started by 200 persons and completed by 46. From these 46 we excluded 4 people who spent less than 10 minutes (TIME) on the questionnaire since we assumed that to be the minimum time to provide meaningful answers. These 42 persons and their answers to the 36 questions establish the sample for our statistical analysis below. Altogether, 1512 answers are recorded in the sample. 65% of the participants had more than three years experience in process modeling.

3.2 Hypothetical Relations between Factors and Understandability

Before conducting the statistical analysis we make hypothetical connections between the different variables explicit. In particular, we identify hypotheses related to personal factors, model factors, and content factors:

P1. A higher PSCORE of participants should be connected with higher values in THEORY, DURATION, INTENSITY, and TIME.

M1. A higher MSCORE of models should be associated with lower values in SIZE, DIAMETER, TOKEN SPLIT, and HETEROGENEITY since these metrics might indicate that the model is easier to comprehend.

M2. A higher MSCORE of models should be connected with higher values in STRUCTUREDNESS, SEPARABILITY, and SOUND since these metrics might be associated with models that are easier to comprehend.

C1. A higher sum of CORRECTANSWER should be connected with abstract labels (value of 0 in TEXT), basically our questions refer to structural properties of the model.

C2. A CORRECTANSWER (value of 1) should be connected with a lower value in TEXTLENGTH, since it becomes harder to match the elements mentioned in the question with the elements in the graphical model.

In the following section we will use statistical methods to assess these hypotheses.

4 Results

In this section we present the results of our survey. First, we discuss the distribution of personal, model, and content factors as well as their correlation with the corresponding score. Then, we use logistic regression to gain insight into the relative importance of the different factors for predicting whether a question would be answered correctly.

4.1 Personal Factors

Figure 3 gives an overview of the PSCORE that the different participants achieved in the survey. The mean value was 25.21 for 36 questions which means that on average 70% of the questions were answered correctly. The best participant had 34 questions correct and two questions wrong. The correlation with some variables of **P1**, i.e. DURATION, INTENSITY, and TIME, was weakly positive, but not significant. In contrast to that, the Pearson correlation between PSCORE and THEORY was positive and significant (p=0.01) with a value of 0.491. This might indicate that theoretical process modeling knowledge helps to answer questions correctly. Furthermore, THEORY was positively and significantly (p=0.01) correlated with INTENSITY of process modeling experience (0.438).

4.2 Model Factors

Figure 4 gives an overview of the MSCORE that the different participants achieved per model. The mean percentage was 70% across the models. The model with the lowest MSCORE had on average 60% correct answers. This model had loops and parallel execution paths. From the variables mentioned in **M1** and **M2** only SEPARABILITY had a significant correlation according to Spearman with MSCORE of 0.886 (p=0.019). This strongly confirms the hypothetical impact direction of

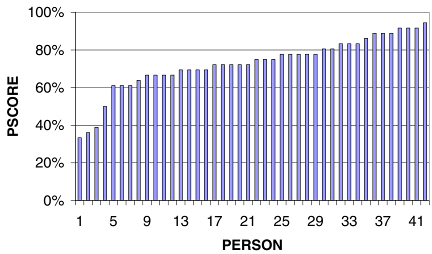


Fig. 3. `pscore` of participants

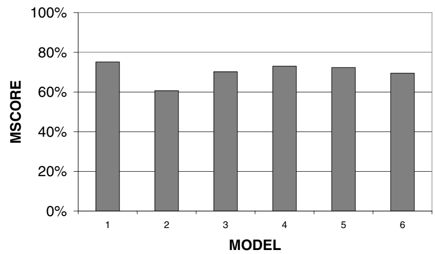


Fig. 4. `mscore` of model

M2. The other variables showed a direction of correlation as expected, but without a sufficient significance. As an exception, structuredness had zero correlation in our sample.

4.3 Content Factors

Table 1 gives an overview of the sum of CORRECTANSWER disaggregated by TEXT. This table directly relates to hypothesis **C1**. The difference between both questionnaire types is quite narrow, such that the hypothesis cannot be confirmed. Furthermore, we calculated the correlation between CORRECTANSWER and TEXTLENGTH. The Pearson correlation coefficient of -0.836 on a significance level above 0.01 supports the hypothesis. Apparently, it becomes harder to match the elements mentioned in the question with the elements in the graphical model when the text labels get larger.

Table 1. Sum of CORRECTANSWER for the two model variants (TEXT=0 is abstract letters and TEXT=1 textual labels

TEXT	correct	all	%
0	608	828	73.43%
1	493	684	72.08%

4.4 Prediction of Correct Answers

Up to now, we have studied bivariate correlations between different hypothetical factors and different score variables as proxies for understandability. In this section we investigate the combined capability of the factors to explain the variance of the dependent variable CORRECTANSWER. This variable captures whether a single question related to a model was answered correctly by a participant. As the dependent variable is binary, we use a logistic regression (logit) model. The idea of a logit model is to model the probability of a binary event by its odds, i.e., the ratio of event probability divided by non-event probability. These odds are defined as $logit(p_i) = ln(\frac{p_i}{1-p_i}) = B_0 + B_1 x_{1,i} + \ldots + B_k x_{k,i}$ for k input variables and i observations, i.e. EPC i in our context. From this follows that

$$p_i = \frac{e^{B_0+B_1x_{1,i}+\ldots+B_kx_{k,i}}}{1+e^{B_0+B_1x_{1,i}+\ldots+B_kx_{k,i}}}$$

The relationship between input and dependent variables is represented by an S-shaped curve of the logistic function that converges to 0 for $-\infty$ and to 1 for ∞. The cut value of 0.5 defines whether event or non-event is predicted. $Exp(B_k)$ gives the multiplicative change of the odds if the input variable B_k is increased by one unit, i.e. $Exp(B_k) > 1$ increases and $Exp(B_k) < 1$ decreases error probability.

The significance of the overall model is assessed by the help of two statistics. Firstly, the *Hosmer & Lemeshow* Test should be greater than 5% to indicate a good fit based on the difference between observed and predicted frequencies [33]. Secondly, *Nagelkerke's* R^2 ranging from 0 to 1 serves as a coefficient of determination indicating which fraction of the variability is explained [34]. Furthermore, each estimated coefficient of the logit model is tested using the *Wald* statistic, for being significantly different from zero. The significance should be less than 5%. We calculate the logistic regression model based on a stepwise introduction of those variables that provide the greatest increase in likelihood. For more details on logistic regression, see [33].

Figure 5 shows the result of the logistic regression estimation. The best model is derived in step 4 since it covers the largest set of variables such that all have a significance in the *Wald* statistic better than 5%. The *Nagelkerke's* R^2 for this

Variables in the Equation[f]

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	THEORY	.148	.009	300.694	1	.000	1.160
Step 2[b]	SEPARABILITY	1.012	.384	6.966	1	.008	2.752
	THEORY	.121	.013	86.815	1	.000	1.129
Step 3[c]	TEXTLENGTH	-.001	.000	2.831	1	.092	.999
	SEPARABILITY	1.088	.387	7.908	1	.005	2.969
	THEORY	.133	.015	80.452	1	.000	1.142
Step 4[d]	TEXTLENGTH	-.001	.000	3.946	1	.047	.999
	SEPARABILITY	.935	.396	5.582	1	.018	2.548
	THEORY	.120	.016	52.634	1	.000	1.127
	DURATION	.006	.003	3.073	1	.080	1.006
Step 5[e]	TEXTLENGTH	-.001	.000	2.815	1	.093	.999
	DIAMETER	-.015	.010	2.493	1	.114	.985
	SEPARABILITY	1.185	.428	7.677	1	.006	3.270
	THEORY	.140	.021	45.103	1	.000	1.150
	DURATION	.008	.004	4.601	1	.032	1.008

a. Variable(s) entered on step 1: THEORY

b. Variable(s) entered on step 2: SEPARABILITY

c. Variable(s) entered on step 3: TEXTLENGTH

d. Variable(s) entered on step 4: DURATION

e. Variable(s) entered on step 5: DIAMETER

f. Stepwise procedure stopped because removing the least significant variable result in a previously fitted model.

Fig. 5. Logistic regression function estimated for the sample

step 4 model is 0.293 indicating that quite a considerable share of the overall variance can be explained. Still, the *Hosmer & Lemeshow* Test is below 5% which signals that there seem to be other factors that are not covered by the model. The model includes four variables, namely THEORY, SEPARABILITY, TEXTLENGTH, and DURATION. While TEXTLENGTH appears to have a negative effect on CORRECTANSWER as expected by **C2**, the other variables have a positive impact on the correctness of answering a question. This confirms the hypotheses **P1** and **M2**. It is interesting to note that the step 4 model includes factors that are related to all three influences that we identified, i.e. personal, model, and content factors with the model factor SEPARABILITY having the greatest relative impact.

5 Conclusions

In this paper we have used an online questionnaire to acquire insight into the empirical connection between different influential factors and process model understandability. In particular, we focused on several personal, model, and content related impact factors. By the help of a correlation analysis we found that the personal factor THEORY (0.491), the model factor SEPARABILITY (0.886 Spearman), and the content factor TEXTLENGTH (-.836) were significantly correlated with the different proxies for understandability. A multi-variate logistic regression analysis confirmed the importance and impact direction of these three variables, and also included DURATION in the statistical model. Altogether, the survey supports the hypothesis that personal, model, and content related factors influence the understandability of business process models. This has strong implications for business process modeling initiatives in organizations. First, there is apparently a need for guidelines that lead to understandable process models in terms of structure and text labels. Second, there is a need for training since experienced modelers perform better in understanding.

The research design used in this paper has some limitations. Firstly, we only investigated three categories of influence factors of understandability. Future research will have to analyze other categories as well. Given the considerable number of impact factors that might be important ([9] mention at least six) it will be difficult to study the relative importance of these factors. Secondly, our experiment covered only a limited set of six models and a limited set of 42 participants. Future surveys need to have more models tested by each participant for comprehending the relative importance of the different model metrics. This implies challenges with motivation and fatigue. Finally, more participants are needed to clearly identify which personal factors have the greatest impact on understanding a process model.

Acknowledgement

The author would like to thank Wolfgang Hammel, Gerald Schneider, Yvonne Tima, and Maria Weiler for their help in designing the questionnaire and conducting the experiment.

References

1. Hoppenbrouwers, S.S., Proper, H.E., van der Weide, T.: A Fundamental View on the Process of Conceptual Modeling. In: Delcambre, L.M.L., Kop, C., Mayr, H.C., Mylopoulos, J., Pastor, Ó. (eds.) ER 2005. LNCS, vol. 3716, pp. 128–143. Springer, Heidelberg (2005)
2. Becker, J., Rosemann, M.: Guidelines of Business Process Modeling. In: van der Aalst, W.M.P., Desel, J., Oberweis, A. (eds.) Business Process Management. LNCS, vol. 1806, pp. 30–49. Springer, Heidelberg (2000)
3. Moody, D.: Theoretical and practical issues in evaluating the quality of conceptual models: current state and future directions. Data & Knowl. Eng. 55, 243–276 (2005)
4. Davies, I., Green, P., Rosemann, M., Indulska, M., Gallo, S.: How do practitioners use conceptual modeling in practice? Data & Knowl. Eng. 58, 358–380 (2006)
5. Mendling, J., Moser, M., Neumann, G., Verbeek, H., Dongen, B., Aalst, W.: Faulty EPCs in the SAP Reference Model. In: Dustdar, S., Fiadeiro, J.L., Sheth, A.P. (eds.) BPM 2006. LNCS, vol. 4102, pp. 451–457. Springer, Heidelberg (2006)
6. Mendling, J., Verbeek, H., Dongen, B., Aalst, W., Neumann, G.: Detection and Prediction of Errors in EPCs of the SAP Reference Model. In: Data & Knowl (accepted, 2007)
7. Simon, H.: Sciences of the Artificial., 3rd edn. The MIT Press, Cambridge (1996)
8. Mendling, J.: Detection and Prediction of Errors in EPC Business Process Models. PhD thesis, Vienna University of Economics and Business Administration (2007)
9. Mendling, J., Reijers, H., Cardoso, J.: What makes process models understandable. In: Alonso, G., Dadam, P., Rosemann, M. (eds.) BPM 2007. LNCS, vol. 4714, pp. 48–63. Springer, Heidelberg (2007)
10. Lindland, O., Sindre, G., Sølvberg, A.: Understanding quality in conceptual modeling. IEEE Software 11, 42–49 (1994)
11. Krogstie, J., Sindre, G., Jørgensen, H.: Process models representing knowledge for action: A revised quality framework. Europ. J. of Inf. Systems 15, 91–102 (2006)
12. Moody, D., Sindre, G., Brasethvik, T., Sølvberg, A.: Evaluating the Quality of Process Models: Empirical Testing of a Quality Framework. In: Spaccapietra, S., March, S.T., Kambayashi, Y. (eds.) ER 2002. LNCS, vol. 2503, pp. 380–396. Springer, Heidelberg (2002)
13. ISO: Information technology - software product evaluation - quality characteristics and guide lines for their use. ISO/IEC IS 9126 (1991)
14. Güceglioglu, A.S., Demirörs, O.: Using software quality characteristics to measure business process quality. In: van der Aalst, W.M.P., Benatallah, B., Casati, F., Curbera, F. (eds.) BPM 2005. LNCS, vol. 3649, pp. 374–379. Springer, Heidelberg (2005)
15. Gemino, A., Wand, Y.: Evaluating modeling techniques based on models of learning. Commun. ACM 46, 79–84 (2003)
16. Green, T., Petre, M.: Usability analysis of visual programming environments: A 'cognitive dimensions' framework. J. Vis. Lang. Comput. 7, 131–174 (1996)
17. Lee, G., Yoon, J.M.: An empirical study on the complexity metrics of petri nets. Microelectronics and Reliability 32, 323–329 (1992)
18. Nissen, M.: Redesigning reengineering through measurement-driven inference. MIS Quarterly 22, 509–534 (1998)
19. Morasca, S.: Measuring attributes of concurrent software specifications in petri nets. In: METRICS 1999, pp. 100–110 (1999)

20. Reijers, H., Vanderfeesten, I.: Cohesion and coupling metrics for workflow process design. In: Desel, J., Pernici, B., Weske, M. (eds.) BPM 2004. LNCS, vol. 3080, pp. 290–305. Springer, Heidelberg (2004)
21. Cardoso, J.: Evaluating Workflows and Web Process Complexity. In: Workflow Handbook 2005, pp. 284–290 (2005)
22. Balasubramanian, S., Gupta, M.: Structural metrics for goal based business process design and evaluation. Business Process Management Journal 11, 680–694 (2005)
23. Canfora, G., García, F., Piattini, M., Ruiz, F., Visaggio, C.: A family of experiments to validate metrics for software process models. Journal of Systems and Software 77, 113–129 (2005)
24. Aguilar, E.R., Ruiz, F., García, F., Piattini, M.: Towards a Suite of Metrics for Business Process Models in BPMN. In: Manolopoulos, Y., Filipe, J., Constantopoulos, P., Cordeiro, J. (eds.) ICEIS 2006, Proceedings (III), pp. 440–443 (2006)
25. Laue, R., Gruhn, V.: Complexity metrics for business process models. In: Abramowicz, W., Mayr, H.C. (eds.) BIS 2006, Proceedings. LNI, vol. 85, pp. 1–12 (2006)
26. Cardoso, J.: Process control-flow complexity metric: An empirical validation. In: IEEE SCC 2006, Proceedings, pp. 167–173 (2006)
27. Rosemann, M., Recker, J., Indulska, M., Green, P.: A Study of the Evolution of the Representational Capabilities of Process Modeling Grammars. In: Dubois, E., Pohl, K. (eds.) CAiSE 2006. LNCS, vol. 4001, pp. 447–461. Springer, Heidelberg (2006)
28. Agarwal, R., Sinha, A.: Object-oriented modeling with UML: A study of developers perceptions. Communications of the ACM 46, 248–256 (2003)
29. Sarshar, K., Loos, P.: Comparing the Control-Flow of EPC and Petri Net from the End-User Perspective. In: van der Aalst, W.M.P., Benatallah, B., Casati, F., Curbera, F. (eds.) BPM 2005. LNCS, vol. 3649, pp. 434–439. Springer, Heidelberg (2005)
30. Lange, C., Chaudron, M.: An experimental investigation. In: Osterweil, L., Rombach, H., Soffa, M. (eds.) ICSE 2006, Proceedings., pp. 401–411 (2006)
31. Mendling, J., Aalst, W.: Towards EPC Semantics based on State and Context. In: Nüttgens, M., Rump, F.J., Mendling, J. (eds.) EPK 2006, Proc., pp. 25–48 (2006)
32. Mendling, J., Aalst, W.: Formalization and Verification of EPCs with OR-Joins Based on State and Context. In: Krogstie, J., Opdahl, A., Sindre, G. (eds.) CAiSE 2007 and WES 2007. LNCS, vol. 4495, pp. 439–453. Springer, Heidelberg (2007)
33. Hosmer, D., Lemeshow, S.: Applied Logistic Regression, 2nd edn. (2000)
34. Nagelkerke, N.: A note on a general definition of the coefficient of determination. Biometrika 78, 691–692 (1991)

Concepts for Modeling Hybrid Products in the Construction Industry

Daniel Weiss, Joerg Leukel, and Stefan Kirn

University of Hohenheim, Information Systems II, Schwerzstr. 35,
70599 Stuttgart, Germany
{daniel.weiss,joerg.leukel,stefan.kirn}@uni-hohenheim.de

Abstract. Methods for modeling products need to be extended or integrated with other methods when moving from products, i.e., tangible goods, to hybrid products. This paper investigates such modeling methods from a method-engineering perspective. In particular, we consider the construction industry which is increasingly subject of hybrid product strategies as a mean of differentiation. The contribution is that we (1) reconstruct core modeling concepts from six modeling methods and (2) integrate those into a meta model. The analysis shows that the so called result dimension is still dominating, thus what hybrid products provide whereas both the process and resource dimension lack attention and dedicated modeling concepts.

Keywords: Hybrid Products, Meta Modeling, Modeling Methods, Product Models, Product Data Management, Reference Models.

1 Introduction

Product Data Management (PDM) offers a comprehensive set of integrated methods for modeling products ranging from parts and components to highly complex products. These methods are widely used in practice and implemented in respective information systems. The methods, however have to face the *problem* of hybrid products. Such products represent integrated bundles of both tangible goods and services, and therefore require considering the resource, process and result dimensions. Hybrid products can exemplarily be studied in the construction industry. Due to increasing price competition in this industry, hybrid products as value-adding solutions to specific customer requirements, have attracted many firms as a mean of differentiation. In addition, this industry has yielded specific modeling methods for products which need to be studied whether they can be extended towards hybrid products.

This paper *investigates* methods for modeling products and services. Such methods have emerged independently and successively from each other and thus differ not only in their underlying design paradigms but also with regard to process model, modeling concepts and notation. We take a *method engineering perspective* which allows for (1) conducting a systematic analysis of syntactically diverse modeling methods and (2) constructing a domain-specific modeling method for hybrid products. The contribution

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 154–164, 2008.

is that we (1) reconstruct core modeling concepts from six modeling methods and (2) integrate those into a meta model. This is made possible by language-based meta modeling as the underlying research method [1]. This research method requires the following four steps: (1) definition of the meta model language, (2) specification of the meta models of each modeling method, (3) identification of common and different modeling concepts and (4) specification of an integrated meta model.

The present work contributes to a research framework which concerns *logistics systems under customization*. Logistics systems provide services which transform goods with regard to location, time, and quantity. The goal of this research framework is to make logistics massively customizable by means of information systems. Customization is a major trend [2] and is in particular observable in the construction industry. Customers demand not single, standardized products which can be bought off-the-shelf, but complex ones which (1) are tailored to their specific needs (e.g., usage and location of the building, preferences regarding architecture, style, interior etc.) and (2) necessarily require services such as consulting, planning, supervision, and building activities. Such customized hybrid products impose also new requirements on logistics in the construction industry. The current work addresses hybrid products which are subject of a set of inter-related logistics services. We study the means how firms in the supply chain of the construction industry model and thus describe such products.

The remainder of this paper is structured as follows. Section 2 reviews related work. In section 3, we design our research approach. In section 4, we present modeling methods for products and services which will be analyzed. In section 5, we perform the model synthesis. The final section draws conclusion and outlines future work.

2 Related Work

The related work can be grouped into product and service modeling respectively.

2.1 Product Modeling

Product modeling is an essential task of product data management. The resulting models define product data in such a way that respective information systems can be designed and product data can be exchanged. The concept of an integrated product model is closely connected with the standard ISO 10303 (STEP standard for the exchange of product model data) [3].

Technical product models represent physical product properties and are described by for instance, geometry, kinematics, flow, and deformation models. These models focus on the product structure and features. STEP allows creating such models by means of data models which can be specified using the data definition language EXPRESS.

In contrast to technical product models, *commercial product models* are used in enterprise information systems such as ERP systems. They contain product data for the functional business areas such as procurement, manufacturing, and distribution. These models can be described by means of standardized data exchange formats and product classification systems. The latter are also known as product ontologies and can, for instance, be specified using ontology languages of the Semantic Web [4].

Regarding hybrid products, the above mentioned methods do not provide concepts for modeling original aspects of hybrid products [5]. Very often, such methods are used for describing services or even hybrid products which results in product models that do not reflect essential characteristics of services or hybrid products (e.g. in [6]).

In terms of comparative studies of modeling methods, approaches based on requirements engineering dominate which compare methods on the basis of empirical or theoretically derived requirements (e.g., [7]).

2.1 Service Modeling

Service modeling is an essential task of service engineering. Service engineering refers to systematically developing new services based on a methodological approach, similarly to the development of products. A service model is the representation of the service to be provided.

Services can be characterized by their constituent properties immateriality and customer integration (so called contact services). Thus, for contact services, the customer interaction has to be part of the service model. A fundamental difference to product modeling gets obvious: Service modeling does not only address the result but primarily the process. Consequently, many service modeling approaches are based on general process modeling methods:

- Service Blue Printing is an early method for the graphical representation of service processes [8]. It is used for service development and does not make a direct contribution to the service representation in information systems. This also applies to extensions of this method such as [9].
- Grieble et al. [10] distinguish three service dimensions: resource, process, and result. For the process dimension, they propose to use EPC method (event driven process chains).
- Schneider and Thomas suggest an extension of the EPC method due covering the customer integration [11].
- In contrast to that, [12] adopt conventional PDM product modeling methods for service modeling and show this approach for education and training services.

Modeling methods from service engineering still suffer from diverse deficits. In particular, the development, deployment, and validation of a universal modeling method is hindered by the fact that real service models differ significantly depending of the respective service industry (e.g., financial vs. training services). These observations suggest that service models are more domain-specific and can only partially be specified using a domain-independent method.

3 Research Approach

In this section, we design our research approach based on meta modeling.

3.1 Meta Modeling Process

The diversity and heterogeneity of approaches for modeling product and services as identified in the previous section provide an obstacle for the move from products and services modeling to hybrid product modeling.

To compare syntactically differing modeling methods, we employ language-based meta modeling. This allows for capturung a method's concepts (analysis) and extracting the constitutive concepts for modeling hybrid products (synthesis). Meta modeling as a research method for comparing methods can be found, e.g., in [13]. This approach requires the following four steps:

1. Determination of the meta model language: Requirements for a suitable meta model language are semantic expressiveness and determination. Semantic expressiveness describes how precise and differentiated the language can be used. Determination requires the absence of design freedom. Situations should be avoided in which different, though semantically equivalent modeling alternatives can be chosen for representing the same concept.
2. Specification of meta models: For each modeling method, the meta model has to be specified. The terminology of the original has to be kept. Generally, three cases can be distinguished: (a) Modeling method with an explicit meta model: The meta model can be taken and, if necessary, converted to the chosen meta model language. (b) Modeling method with no explicit meta model: The meta model needs to be reconstructed. (c) Reference data models: At first, relevant concepts have to be typecasted. Afterwards, these can be extracted and compared to the modeling concepts of 2a and 2b.
3. Identification of common and different concepts: Due to a common language, the specified meta models are syntactically comparable. Nevertheless, when comparing conceptual models, three sorts of model conflicts need to be considered and solved: type, naming and structural conflicts. Naming and structural conflicts are caused by a heterogeneous usage of the domain language in the conceptual model. Type conflicts, however, can be traced to a different representation of reality on the level of the modeling language.
4. Specification of an integrated meta model: In this step, the identified *relevant* modeling concepts will be consolidated and integrated into a single meta model for hybrid products.

3.2 Meta Model Language and Modeling Conventions

We use the UML diagram type *class diagram* as meta model language. This language fulfills the requirement of semantic expressiveness, since it provides a large number of modeling concepts such as classes, attributes, generalization / specialization, aggregation, association, and association classes. It is worth noting that class diagrams do not fully comply with the requirement of determination [14]. Therefore, further modeling restrictions on the degree of freedom are necessary (modeling conventions):

1. Vertically, classes will be arranged according to structure. Specialized classes are placed below more general classes.
2. Horizontally, classes are arranged from left to right whether they depend on the existence of other classes

These rules materialize in the layout shown in figure 1.

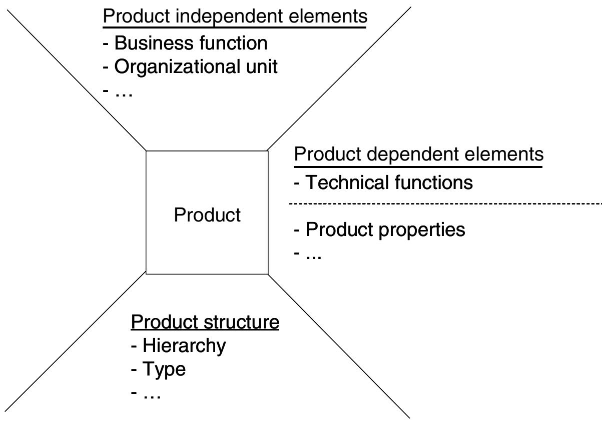


Fig. 1. Layout of Meta Models

The product structure is placed below the central product class. Product independent elements such as organizational unit related to products can be found in the upper part of the model. Product dependent elements such as functions and properties are subject of the right hand area.

As already mentioned, the original terminology of each modeling methods will be retained as far as possible. To increase clarity, attribute lists will not be included in the class symbols. Due to diverse focus of some modeling methods, the respective meta models can not be covered completely. This is especially valid for those methods that provide highly detailed concepts for areas such as product prices.

4 Analysis

In this section 4, we present modeling methods for products and services which will be analyzed.

4.1 Overview

Subject of investigation are six selected modeling methods (see table 1) which represent the most relevant methods due to maturity and adoption by industry. With regard to the method engineering perspective, the methods form two groups:

- Conceptual modeling methods: For each method the meta model needs to be adopted, adapted, or reconstructed (here: ARIS and ISO 10303).
- Reference data models: The model elements have to be typecasted respectively conceptualized at first. Eventually, they can be used for comparison (here:: ISO 13584, Y-CIM reference models as well as GAEB DA XML 3.0 and bau:class).

Table 1. Product Modeling Methods

Name	Reference	Type	Notation	Meta model construction
ARIS	[15]	product model language	custom	adoption
ISO 10303	[3]	data definition language	custom	reconstruction
ISO 13584	[16]	reference data model	EXPRESS	
Y-CIM	[17]	reference data model	ERM	typecasting and
GAEB DA XML 3.0	[18]	reference data model	XML Schema	reconstruction
bau:class	[19]	reference data model	Database schema	

While ISO 10303 and ISO 13584 address the modelling of products, ARIS and Y-CIM are not limited to a tangible goods. GAEB DA XML 3.0 and bau:class represent specific modeling methods used in the German construction industry.

4.2 ARIS

ARIS provides concepts for modeling goods and services [15], which are essential inputs and outputs for business processes. Following other industrial approaches, products in ARIS are hierarchically structured. Furthermore, products can be attributed with a type of costs and cost rates. In [15], the product meta model is already modeled by means of class diagrams. Compared to other PDM-oriented product modeling methods, ARIS focuses product structures and refers for a detailed representation of product features to other concepts, such as industrial product catalogs [17].

4.3 ISO 10303

ISO 10303 [3] provides concepts for modeling products by means of the EXPRESS language [20]. EXPRESS combines concepts of relational and object-oriented modeling and is supported by the semi-formal method EXPRESS-G. The concepts include entities, which represent classes of objects with some same properties, and typed relationships between entities. Attributes describe entities and can be atomic or structured. Schemes summarize several entities that furthermore can be referenced from other, external schemes.

4.4 ISO 13584

ISO 13584 concerns the standardization of product classes and technical product properties. For this purpose, it defines a reference data model [16]. As a complementary standard to ISO 10303, it uses both EXPRESS and EXPRESS-G for specifying this model. An UML-based representation of this model can be found in [21].

4.5 Y-CIM

The Y-CIM reference model for industrial business processes is used for structuring and describing business and technical tasks in industrial enterprises. The reference model is based on ARIS family of methods. However, unlike ARIS, product modeling is subject of the data view [17], because it is based on an earlier version of ARIS which does not provide a specific product view. Products are referred to as parts, which can be structured hierarchically and described by means of properties. Further concepts are principles and technical features as well as functional structures and relationships between properties.

4.6 GAEB DA XML 3.0

GAEB DA XML 3.0 [26] is a standard for the exchange of data in the German construction industry. The respective reference data model is specified using XML Schema. This model covers all phases of construction from planning to execution. Products in the construction industry are characterized by pre-defined attributes and textual descriptions. A product description includes an explanation of work and optional documents. The work can be hierarchically structured and divided in so called lots of number (e.g., buildings 1, 2, 3, etc.) and content (such as earthwork, carpentry etc.).

4.7 Bau:Class

bau:class is a classification system (or taxonomy) for products and services in the German construction industry [19]. The classification includes about 5,000 definitions. Products and services classes are described by means of property lists. Such classes may represent complex products though such a product structure is not explicitly described.

5 Synthesis

In this section, we report about the results of the synthesis of all meta models.

5.1 Model Elements and Scope

The method analysis and synthesis calls for a systematic comparison of the modeling concepts found in the respective methods. Therefore, the original language has to be retained to compare the semantic. The comparison includes only those elements that are explicitly contained in the method's meta model by means of classes or associations. The attribute level is not considered here. Table 2 shows which elements the methods contain and how they are entitled.

Reviewing the table above, the following interim conclusions can be drawn:

- ARIS distinguishes products in the sense of products and services and embeds product modeling in enterprise modeling.
- In Y-CIM, relations to manufacturing processes can be described by means of technical functions and solution principles. These two concepts are missing in all other methods.
- All considered methods describe products by means of properties. ISO 13584 presents the most comprehensive property concept and allows dependencies between properties.

Table 2. Comparison of Meta Model Elements and Terminology

ARIS	ISO 10303	ISO 13584	Y-CIM	GAEB DA XML 3.0	bau:class
Product					
Product	Entity	Product class	Part	Building product	Product
Good Service Information service Miscellaneous service					
Product structure					
Product structure	Relation		Part structure	Product structure	
	Pattern, Relation			Lot	
					Subject area
Product-dependent elements					
Attribute	Attribute	Property	Part property	Attribute	Description property Property list
				Product description	
	Synonym				Keyword
			Technical function		
		Domain Value			Characteristic
Product-independent elements					
Type of costs Cost rate					
			Solution principle		
		Supplier, Identification			

– ISO 13584 and bau:class consider only atomic products, whereas the other methods allow for modeling complex products by means of hierarchical structures.
– Product bundles can only be described in ISO 10303 and GAEB.

5.2 Integrated Meta Model

The aim of an integrated meta model is to consolidate the relevant concepts for modeling hybrid products. It is not just a superset of the individual meta models.

For integrating the identified concepts and systematically constructing the model, we adopt the object-type approach which provides the four operations subsumption, subordination, composition, and reduction (a detailed explanation can be found in [2]).

The meta model shown in figure 2 is based on the results of the previous analysis and comparison:

- Products are distinguished in tangible goods, services, and hybrid products.
- Product characteristics are described by means of properties. The property model is based on ISO 13584. Textual descriptions, as they are made explicit in GAEB, are represented by properties.
- According to Y-CIM and to take into account the customer integration, the concept of customer function has been considered. These functions can have a hierarchical structure.

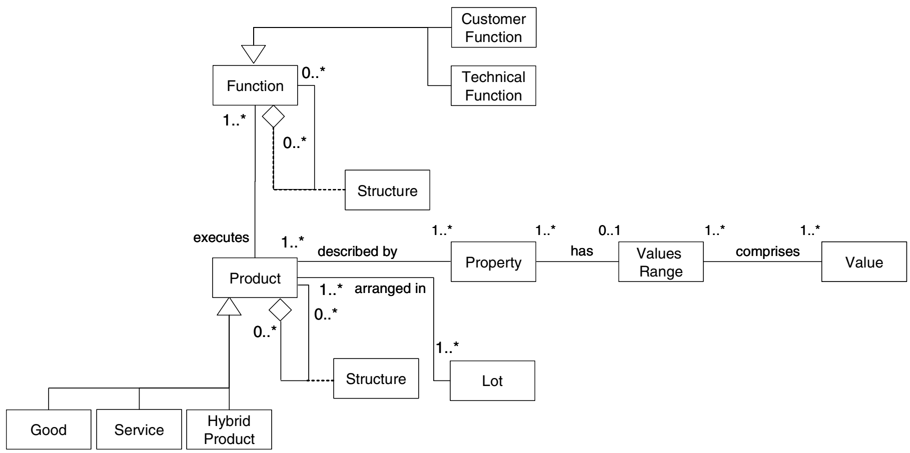


Fig. 2. Integrated Meta Model

6 Conclusions

In this paper, methods for modeling products and services were investigated. In doing so the question should be answered, to what extent the considered methods provide concepts for modeling hybrid products. For this purpose, we adopted a method engineering perspective, which allowed by means of language-based meta modeling a systematic analysis of syntactically diverse modeling methods. As a result an integrated meta model was constructed.

The contribution is that we (1) reconstruct core modeling concepts from six modeling methods and (2) integrate those into a meta model. The analysis shows that the so called result dimension is still dominating, thus what hybrid products provide whereas both the process and resource dimension lack attention and dedicated modeling concepts.

With regard to the research framework of *logistics systems under customization*, we studied the means how firms in supply chains of the construction industry could model and thus describe hybrid products. The analysis, however, indicates that such means are hardly available. The conclusion is that interactions between customers and suppliers with the purpose of matching customer requirements with supplier capabilities require additional modeling concepts and more suitable methods. Thus the current methods in the construction industry do not cater for individualization whereas they focus rather standardized, less complex products.

The considered methods originate predominantly from product modeling in the sense of goods, even if they sometimes allow a differentiation between goods and services. It can be stated, however, that modeling concepts for the customer integration in particular, hardly exist. This way the methods address the result dimension only.

Relating to the process dimension of hybrid products, dedicated concepts are hardly recognizable. The example of the construction industry presents a pragmatic approach for representing hybrid products. The modeling approach is based on just a few constructs, the formal semantics is limited.

Since the current meta model focuses primarily result-oriented aspects, future work has to aim at integrating the resource and process dimension, which are required for a clear understanding of hybrid product modeling.

Acknowledgment

This work was supported by the German Ministry for Research and Technology (BMBF) under grant no. 01FD0667-01DF0671.

References

1. Karagiannis, D., Kuehn, H.: Metamodelling Platforms. In: Bauknecht, K., Min Tjoa, A., Quirchmayer, G. (eds.) DEXA 2002. LNCS, vol. 2455, pp. 451–464. Springer, Heidelberg (2002)
2. Coates, J.: Customization promises sharp competitive edge. Research in Technology Management 38, 6–7 (1995)
3. International Organization for Standardization: ISO 10303-11:2004. Industrial automation systems and integration – Product data representation and exchange – Part 1: Overview and fundamental principles. Geneva (1994)
4. Hepp, M.: A Methodology for Deriving OWL Ontologies from Industrial Categorization Standards. International Journal on Semantic Web & Information Systems 2(1), 72–99 (2006)
5. DFG Collavorative Research Project Transregio 29 (SFB/TR 29), http://www.lps.ruhr-uni-bochum.de/tr29/english
6. Lee, T., Shim, J., Lee, H., Lee, S.-g.: A Pragmatic Approach to Model and Exploit the Semantics of Product Information. In: Spaccapietra, S. (ed.) Journal on Data Semantics VII. LNCS, vol. 4244, pp. 242–266. Springer, Heidelberg (2006)
7. Kelkar, O., Leukel, J., Schmitz, V.: Price Modelling in Standards for Electronic Product Catalogs Based on XML. In: Proceedings of the 11th International World Wide Web Conference (WWW 2002), pp. 366–375. ACM, New York (2002)
8. Shostack, L.G.: How to Design a Service. European Journal of Marketing 16, 49–63 (1982)
9. Fließ, S., Kleinaltenkamp, M.: Blueprinting the service company. Managing service processes efficiently. Journal of Business Research 57, 392–404 (2007)
10. Grieble, O., Klein, R., Scheer, A.-W.: Modellbasiertes Dienstleistungsmanagement. IWI Heft 171, Universitaet des Saarlandes (2002)

11. Schneider, K., Thomas, O.: Kundenorientierte Dienstleistungsmodellierung mit Ereignisgesteuerten Prozessketten. In: Nuettgens, M., Rump, F.J. (eds.) EPK 2002, Bamberg, pp. 87–93 (2003)
12. Rex, S., Hettrich, A.: Produktdatenmanagement für Dienstleistungen: am Beispiel von Aus- und Weiterbildungsdienstleistungen. In: Dangelmaier, W. (ed.) Innovationen im E-Business, Paderborn, pp. 479–486 (2003)
13. zur Muehlen, M.: Evaluation of Workflow Management Systems Using Meta Models. In: Proceedings of the 32nd Hawaii International Conference on Systems Sciences (HICSS 1999), IEEE Computer Society, Los Alamitos (1999)
14. Fuentes, J.M., Quintana, V., Llorens, J., Génova, G.: Errors in the UML Metamodel? ACM SIGSOFT Software Engineering Notes 28(6), 3 (2003)
15. Scheer, A.-W.: ARIS - Business Process Modelling, 3rd edn. Springer, Heidelberg (2000)
16. International Organization for Standardization: ISO 13584-1:2001. Industrial automation systems and integration – Parts library – Part 1: Overview and fundamental principles. Geneva (2001)
17. Scheer, A.-W.: Business Process Engineering: Reference Models for Industrial Enterprises. Springer, Berlin (1998)
18. Gemeinsamer Ausschuss Elektronik im Bauwesen: GAEB DA XML, http://www.gaeb.de
19. bau:class: bau:class – Das Klassifikationssystem der Baubranche, http://www.bauclass.org
20. International Organization for Standardization: ISO 10303-11:2004. Industrial automation systems and integration – Product data representation and exchange – Part 11: Description methods: The EXPRESS language reference manual. Geneva (2004)
21. European Committee for Standardization (CEN): CWA 15295. Description of References and Data Models for Classification. Brussels (2005)
22. Holten, R.: Entwicklung von Fuehrungsinformationssystemen. Ein methodenorientierter Ansatz. Diss. Wiesbaden (1999)

Improving Requirements Analysis through Business Process Modelling: A Participative Approach*

Jose Luis de la Vara and Juan Sánchez

Department of Information Systems, Valencia University of Technology, Valencia, Spain
{jdelavara,jsanchez}@dsic.upv.es

Abstract. Although requirements analysis is acknowledged as a critical success factor of information system development for organizations, mistakes are frequent at the requirements stage. Two of these mistakes are the lack of understanding of the business by requirements engineers and the miscommunication between business people and systems analysts. As a result of these problems, information systems may not fulfill organizational needs. To prevent these problems, this paper describes an approach based on business process modeling. The business environment is modeled in the form of BPMN diagrams. The diagrams are validated by end-users and are then analyzed by systems analysts in order to reach an agreement on the effect that the information system will have on the organization. Finally, requirements are specified by means of the description of the business process tasks to be supported by the information system.

Keywords: Requirements analysis, understanding of the business, communication, business process, BPMN.

1 Introduction

Requirements analysis is acknowledged as a critical success factor for software projects [32]. If not properly addressed, requirements can cause a project to fail. Nevertheless, practical experience proves that mistakes are frequent at this stage of information system (IS) development for organizations. Two of these mistakes are the lack of understanding of the business by requirements engineers and the miscommunication between business people and computing people. These problems can hinder business/IT alignment [19][27], thus IS does not fulfill the needs of the organization.

Requirements must be defined in terms of phenomena that occur in the business environment [38]. However, it is common for requirements documentation to be solution-oriented, to not reflect the business environment, or to only consist of a data model in the form of a class or entity-relationship diagram. As a solution, several authors have pointed out the importance of organizational modeling during

* This work has been developed with the support of the Ministry of Education and Science of Spain under the project SESAMO TIN2007-62894 and the program FPU, and cofinanced by FEDER.

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 165–176, 2008.

requirements analysis [3][4][37] and the role that requirements engineers play as business analysts [15][28]. Organizational models depict the structure and behavior of an enterprise and are very useful in helping developers properly understand the business environment and the requirements that an information system must fulfill.

Furthermore, good communication between business people and systems analysts is essential at the requirements stage [17][20][30], however it can be difficult to achieve. There is a gap between the business domain and the computing domain that can cause mismatches between what customers say and what requirements analysts understand. Even worse, what customers think they have said and what requirements analysts think they have understood can be totally different. One reason for this miscommunication is that the requirements models that are used can be hard to understand and difficult to validate by customers because of their lack of computing background. Therefore, models that facilitate communication during requirements analysis should be used.

To prevent these problems, business process modeling is a good solution. Business process modeling has not only been acknowledge as a good means for organizational modeling, but also as a must for information system development [1][12][25]. In addition, business process models facilitate human understanding and communication by sharing a common representational format [8]. Among the several notations for business process modeling currently available is BPMN (Business Process Modeling Notation) [23]. It has been designed to be easily understandable, and it is positioned as the de facto standard for business process modeling.

Although there are several related works that define solutions to the problem of linking organizational modeling and requirements engineering (which will be described in more detail in Section 2), the main contribution of this paper consists in providing a methodological approach for deriving the software functionality from organizational models. The approach allows systems analysts to properly understand and analyze the organization, its needs, and goals in a participative way with business people and end-users.

In short, this work describes a requirements analysis approach based on business process modeling. This approach is the result of a project between the Technical University of Valencia and the company CARE Technologies [5]. The business environment is modeled in the form of BPMN diagrams. The diagrams are validated by end-users and are then analyzed in order to reach an agreement on the effects that an IS will have on the organization. Finally, requirements are specified by means of the description of the business process tasks to be supported by the information system. These well-defined requirements are the input of the subsequent development stages.

The paper is organized as follows: section 2 presents related works; section 3 presents the general description of the approach; sections 3.1, 3.2 and 3.3 describe the stages of the technique in detail; section 4 describes practical experience using the approach; finally, section 5 presents our conclusions and future work.

2 Background and Related Work

This section presents background research: section 2.1 describes BPMN; section 2.2 review organizational modeling and goal-oriented requirements approaches.

2.1 BPMN

We use BPMN for business process modeling because it offers a notation that is understandable by all business process users (process analysts, IS developers, process managers…). Therefore, BPMN provides a standard that fills the gap between business models and their implementation. The notation consists of a diagram, called Business Process Diagram (BPD), whose aim is to provide a means for the development of graphical models of business process operations. A BPD is designed from a set of graphical elements that make diagrams simple to develop and easy-to-understand. The graphical elements are flow objects, connecting objects, swimlanes, and artifacts [23].

Several surveys have evaluated the adequacy of BPMN for business process modeling [22][35][36]. From our point of view, BPMN has three main advantages: it is one of the most expressive notations, it is easy to use and understand, and it has been receiving strong support from both practitioners and vendors. As a result, BPMN is considered to be the de facto standard for business process modeling.

2.2 Related Work

The need for organizational modeling in requirements engineering has been widely acknowledged. Many approaches consider it to be the first step in software development, and some of them use business process modeling. Nevertheless, the use of models and techniques that facilitate the communication with customers is not common. Some of the approaches that have received attention in the last few years are reviewed in this section.

The i* framework [37] is one of the most popular techniques for organizational modeling. It has been used in several requirements engineering and software development methods, such as Tropos [6]. i* is a goal-oriented technique that is focused on the dependencies among the organizational actors. Its models are considered to be strategic because actors are not only interested in achieving their own goals, but are also interested in relationships with other actors. As stated by other authors [3][14], the i* framework has some deficiencies. Its models can be too complex, and they do not support granularity and refinement.

EKD [4] provides a way of analyzing an enterprise by using enterprise modeling. It is composed of a goal model, a business rules model, a concepts model, a business process model, an actor model, a resources model, a technical components model, and a requirements model. In our opinion, EKD may be inconvenient to use because none of its models are standard. Furthermore, it lacks tool support to facilitate the development and maintenance of all its models.

Some approaches use UML models for organizational and business process modeling [13][21]. These approaches use elements that are close to those elements used in software development areas. However, this fact is a drawback because models that are easy to use and understand by computing people tend to be too complex to be validated by customers. In addition, the UML-based approaches do not clearly specify some aspects such as the technology that implements business processes or the relationships among the different organizational views.

The ARIS method [29] provides a framework for developing, optimizing and describing integrated information systems. The architecture is structured in four levels: process engineering, process planning and control, workflow control, and application systems. The aim of the ARIS framework is to describe an information system for supporting business process. Therefore, the core of ARIS is the business process, which is a sequence of activities in an enterprise to generate output for the process customer. A reduction of complexity within enterprise modelling is achieved by using different views. The different views are connected within the process view. The most important specification document for presenting the business process model is the event-driven process chain (EPC) diagram that integrates all possible views. ARIS contains different techniques for modelling: entity relationship (ER), UML class diagram, use case, and EPC, however a concrete description for the use of each model during a specific phase is missing [10].

Another interesting and more recent technique is B-SCP [3], which integrates business strategy, context and process using a requirements engineering notation for each of them. Its purpose is to enable the verification and validation of requirements in terms of alignment with and support for business strategy, as well as with the business processes that support that strategy. B-SCP uses RADs [25] for business process modeling, which is less expressive than BPMN. In addition, it has problems associated with goal decomposition, and we think that its requirements specification should be more precise.

3 Approach for BPMN-Based Requirements Engineering

In this section, we present a precise method to guide the construction of requirement models corresponding to organizational models using BPMN diagrams, according to the ideas introduced in section 1. Figure 1 shows a schematic representation of the activities contained in the proposed method. The approach consists of three stages: organizational modeling, business process analysis, and functional requirements specification. The first stage depicts the current business environment (As-Is), which has some problems to be solved or needs to be fulfilled by the new (or modified) IS. Since the organization will change due to the IS development (To-Be), the changes may have an effect on business processes.

In this first stage, the organization for which the IS is going to be developed is modeled. The information gathered is the following: a glossary, the business events, the business rules, a role model, a process map, and the domain data. Business process diagrams of the organization are created from this information.

During business process analysis the diagrams are analyzed in order to reach an agreement on the effects that the development of the information system may have on the business process. One of these effects is the support or control that the information system will have. The business process elements are labeled according which tasks the system will carry out and which tasks will be manually executed. Changes in the business process can occur.

Finally, functional requirements are specified by means of the description of the business process tasks to be supported by the IS. Every task will have a textual template that describes it. The set of templates will be the starting point for the rest of the development stages.

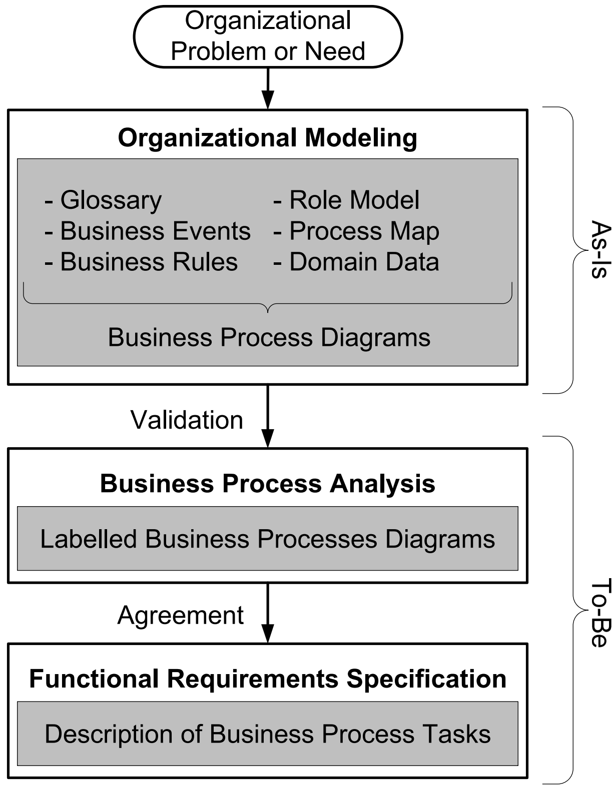


Fig. 1. Approach overview

3.1 Organizational Modeling

To model an organization, the first step is to interview the staff so that the people that play roles in the organization can describe their work. In addition, it is advisable to look through the available documentation related to the organizational activity and the business policies. The purpose of the information collected in the organizational modeling stage is to understand the business environment to be able to model the business process correctly.

The processes are classified into three categories in the process map [16]: management processes, which are the processes related to the management and concern key factors in the long term; core processes, which are linked with product creation and/or service provision; and support processes, which support operative processes and usually concern resources and measures. A glossary is used to define all the organizational concepts unambiguously. Business events are recurrent and significant things that occur while the organization activity goes on and to which the organization must respond. Business rules constrain or define the organizational data and behavior. The different roles and the activities that the roles are responsible for are specified in the role model. The domain data are the entities of the organizational domain that are the input and the output of business process tasks. They are represented by means of a class diagram where there are only classes and the relations among them.

Finally, business processes are modeled from the weaving of all the above information. A business process can be defined as a complete and dynamically coordinated set of collaborative and transactional activities that deliver value to customers [31] and collectively realize a business objective or policy goal [34]. Every business process of the process map has a BPD. BPDs are created from the weaving of all the information gathered, so BPMN graphical elements correspond to this information. The organization is modeled as a pool of every BPD, and the roles are modeled as the lanes. The activities of the roles are modeled as tasks and are included in the BPDs. Events have to be classified as start, intermediate, and final. Each event is included in the BPD where the activity to be triggered is located. Business rules are modeled as gateways, or defined as documentation of the business process tasks, if they cannot be represented graphically. Data objects are included in BPDs later.

The process of organizational modeling is not straightforward and requires a good interaction between the customer and the requirements engineer. Customers must validate the BPDs to guarantee that they properly reflect the organization, and several iterations are usually needed.

3.2 Business Process Analysis

In the second stage, once the organizational modeling is finished, the requirements engineer has enough knowledge of the business to properly understand its activity.

The introduction of an IS in an organization requires business process reengineering and its effect can be assessed before, during or after a process is designed [2]. The new business process (to-be) can be designed from the original one (as-is), from the organizational problem or need, or from the solutions that the IS can provides. The IS will support the required business processes and will be designed in terms of the IS capabilities.

A BPD depicts both the high level requirements specification and the business requirements of the IS. The elements of the business processes are analyzed to establish the effects that the IS will have on them, to determine the elements that will disappear, and to determine the new elements that will be introduced to improve the process. Different alternatives can be proposed, and a solution must be agreed upon with the customer.

When a solution is agreed upon, the business process elements are labeled according to the IS support. Events and gateways can be labeled as follows: “O” (out of the system) if the element will not be part of the system and will not affect it, “IS” (controlled by the system) if the IS will be in charge of the control and execution of the element with no human participation. Apart from these two labels, tasks can be labeled as “U” (executed by a user) if they will be executed by a user that interacts with the IS.

Figure 2 shows the application of business process analysis to the case study of a company that rents apartments for holidays on the coast.

The organization has increased its number of apartments due to growth in the number of customers. Therefore, an IS is needed to facilitate the organization’s activity and increase its efficiency. As an example, the check-out process is used. Its

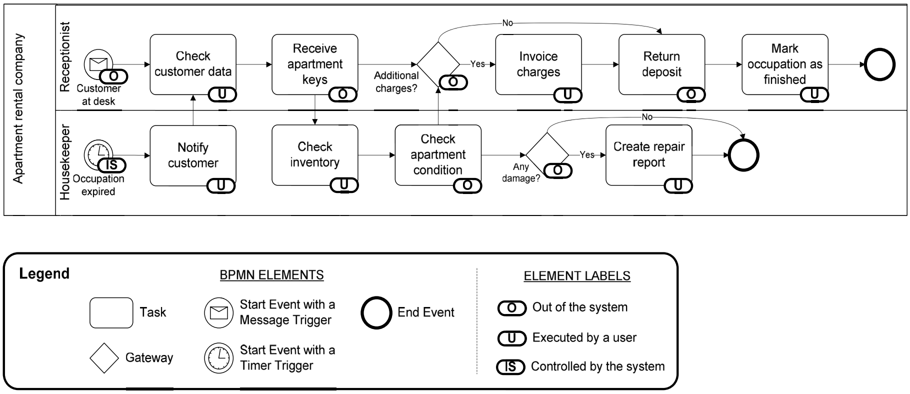


Fig. 2. Labeled check out process of an apartment rental company

purpose is to have the apartments ready to be cleaned so that new customers can move in. The process starts when a customer goes to the reception desk to check out, or when occupation of the apartment has expired and the customer has yet not checked out. In this case, the housekeeper must notify the customer of this fact. Once the receptionist checks the customer data and receives the apartment keys, the housekeeper must check the inventory of the apartment and its condition in order to guarantee that nothing is missing or damaged. If there is something damaged, the housekeeper will create a repair report and the receptionist may charge the customer for existing damages. Finally, the receptionist returns the deposit to the customer and marks the occupation as finished. Data flow (through BPMN data objects) is not shown in the figure for reasons of clarity.

Among several alternatives, Figure 2 shows the solution that was agreed upon with the rental company so that the business process was supported by the IS. The system will be in charge of triggering the occupation expiration. The housekeeper will use the IS when notifying customers, checking the apartment inventory and creating repair reports. Lastly, the receptionist will use the system when checking the customer data, invoicing charges and marking occupations as finished.

3.3 Functional Requirements Specification

In the last stage of the approach, functional requirements are specified from the labeled BPD. For this purpose, business process tasks to be supported by the system (labeled as executed by a user or controlled by the system) are described through a textual template.

The content of the template is based on Constantine's essential use cases [7] and Lauesen´s task & support descriptions [18]. An essential use case is a simplified and generalized form of a use case. It depicts an abstract scenario for one complete and intrinsically useful interaction with a system as understood from the perspective of the users. A task & support description is a way to express what the system actors want to perform, including domain-level information and how a new system could support an activity to solve a problem. Both essential the use cases and the task & support descriptions contain the fewest suppositions about technology.

An example of task template is shown in Table 1. It corresponds to the "Notify customer" task of the check-out process. The template includes the business process to which the task belongs, the name of the task, the role responsible for its execution, the trigger preconditions, and the postconditions of the task, the input and output data and their states, and a specification of the interaction between a user and the IS through user intention and system responsibility.

All the information of the task templates comes from BPDs. The name of the business process and the name of the task are the same as in their BPD. The role is the participant in the business process that is in charge of the task. The triggers correspond to the events with a trigger that precede the task in the business process and have to be part of the IS (controlled by the system), in the same way the preconditions correspond to the gateways that represent decisions and are part of the IS. The postconditions section describes what the change in state of the system will be after the task is completed. The input and output of the task is its data object in the business process. Finally, user intention and system responsibility are defined from the behavior of the participant in charge of the task when executing it and how the system will support it. User intention can also include actions that do not represent interactions with the system. The business rules specified in the template correspond to the business rules defined as the documentation of the task in the organizational modelling stage.

After every business process task has been described, the subsequent development stages will base their work on the task templates to provide the organization with an IS that fits its needs, its structure, and its behavior.

Table 1. Template for the Notify Customer Task

<table>
<tr><td colspan="4">Business Process: Check out</td></tr>
<tr><td colspan="2">Task: Notify Customer</td><td colspan="2">Role: Housekeeper</td></tr>
<tr><td colspan="4">Triggers</td></tr>
<tr><td colspan="4">• Occupation expired</td></tr>
<tr><td colspan="4">Preconditions</td></tr>
<tr><td colspan="4">-</td></tr>
<tr><td colspan="4">Postconditions</td></tr>
<tr><td colspan="4">Updated information stored in System</td></tr>
<tr><td colspan="2">Input</td><td colspan="2">Output</td></tr>
<tr><td>Data Object</td><td>State</td><td>Data Object</td><td>State</td></tr>
<tr><td>Occupation</td><td>Expired</td><td>Occupation</td><td>Expiration Notified</td></tr>
<tr><td>Customer</td><td>-</td><td>-</td><td>-</td></tr>
<tr><td colspan="2">User intention</td><td colspan="2">System responsibility</td></tr>
<tr><td colspan="2"></td><td colspan="2">1. Show the apartments whose occupation has expired</td></tr>
<tr><td colspan="2">2. Select an apartment</td><td colspan="2"></td></tr>
<tr><td colspan="2"></td><td colspan="2">3. Show the occupation and the customer data</td></tr>
<tr><td colspan="2">4. Phone the customer</td><td colspan="2"></td></tr>
<tr><td colspan="2">5. Notify the customer of the occupation expiration</td><td colspan="2"></td></tr>
<tr><td colspan="2">5. Mark the occupation expiration as notified</td><td colspan="2"></td></tr>
<tr><td colspan="2"></td><td colspan="2">6. Record the notification of occupation expiration</td></tr>
<tr><td colspan="4">Business Rules</td></tr>
<tr><td colspan="4">1. Clients are to check out by 10.00 a.m. on the last day of their stay at the latest, and they are obliged to have vacated the apartment by this time.</td></tr>
</table>

4 Practical Experience

The research project presented in this paper has been done in the context of the OO-Method project [24]. OO Method is an industrial model transformation method that relies on a CASE tool to automatically generate complete information systems from object-oriented conceptual models. As stated above, the approach is the result of a project with a company, CARE Technologies. The purpose of the project was to solve problems related to the requirements stage by trying to link the business domain and the software domain.

After analyzing the requirements practices of CARE, we identified problems related to business understanding, purpose analysis of the IS, and communication with end-users. The company uses OO-Method [26], a methodology for automatic software generation based on data conceptual modelling. The conceptual schemas consist mainly of a class diagram that is enriched with functional information about the result of class service execution. Analysts usually use class diagrams as a unique system model, by simply providing a textual description about the requirements and validating them on the class diagram or when the application has been generated.

Although analysts feel comfortable with this technique for information system development, we think it could be improved. Some authors have stated that class diagrams alone might not be appropriate for communicating and verifying requirements, that there are few empirical studies addressing the ability of end-users to understand class models, and that they can be very complex for people that have not been trained in object-oriented modelling [11]. In addition, objects might not be a good way of thinking about a problem domain [33].

The approach has been used in three projects in order to evaluate it and to try to find deficiencies and make improvements. It is currently being used on new projects. The approach has been refined gradually based on the comments of both customers and analysts. The projects developed were a car rental company, the organization of a golf tournament, and the organization of this case study. They are small/medium size projects. CARE had developed software for the companies previously, so both the techniques that they usually use and the approach that we have proposed could be compared.

For the case study (apartment rental company), the introduction of the new IS did not change the business process significantly, and the result was automation rather than reengineering.

We held five meetings to obtain the requirements specification of the case study: two meeting for organizational modelling, one meeting to define the task templates, and one meeting to validate the entire requirements specification. A fifth meeting was held to talk about the experience with end-users and analysts. Each meeting took approximately 2 hours.

As expected, the end-users stated that they could understand and validate the requirements models of the approach more easily than the class diagrams, thus facilitating communication and interaction. They also felt more involved in system development and claimed that they had a more participative attitude.

When asking analysts about the usefulness of the approach, we obtained different opinions. Although all the analysts stated that the approach allowed them to better understand the organizations, the purpose of the IS, and, consequently, the requirements, there were some analysts who did not think the approach could improve their job significantly and would probably not use it.

We do not find these comments about the usefulness of the approach discouraging. When analyzing the opinions more deeply, the analysts that did not think that the approach was as useful were senior analysts who are already very skilled in modelling with OO-Method and interacting with customers. In fact, some of them said that they usually model the systems while the customer describes what the system should do, so they can quickly generate it, validate it, and fix it if needed. However, most of the junior analysts, who have less experience in dealing with customers and, therefore, in understanding what is wanted or needed, considered that the approach could really help them.

We think these results are a reflection of common practices in information system development. Models are only used when they are believed to be useful [9]. In our case, some senior analysts do not think that the approach can accelerate their job, whereas junior analysts think that it can improve their performance.

Finally, this practical experience has some limitations. First, we think the approach has to be used in more projects to draw definite conclusions. Second, the opinions of both the customers and the analysts were obtained by discussing with them informally, so we are now designing a specific form to survey the next projects. Last but not least, we also want to assess the approach by means of experiments with students and software developers from other companies.

5 Conclusions and Future Work

A methodological approach to guide the process of deriving requirement models from organizational models has been presented in this paper. As a main contribution, the approach allows requirement engineers to properly understand the organization and its environment in a participative way with customers. Business people and requirement engineers both share a common language thanks to BPMN and task templates. BPDs are the basis for the customer to validate that the organizational structure and behavior is properly understood so that the requirements engineers can propose solutions for the organization through an IS. This is an important point because the approach implies the involvement of business people, which is another success factor in software development [32].

Requirements analysis is still a stage of software development where mistakes are common. Therefore, it can be the source of problems in subsequent development stages, which can cause an IS not to fulfill the real needs of the organization. Two mistakes which have been detected in practice are the lack of understanding of the business by requirements engineers and the miscommunication between business people and computing people.

As stated above, the approach is the result of a project with a company whose purpose is to link business and software domains. Future work on the project involves the development of a tool that supports the approach. It also involves: the linking of this approach to the OO-Method [26], which is a methodology for automatic software generation based on conceptual modeling, and the introduction of a technique for the analysis of non-functional requirements.

In addition, we want to extend the approach by introducing information about the user interface in the task templates in order to derive an abstract description of the interaction between the users and the IS.

References

[1] Alexander, I., Bider, I., Regev, G.: Workshop on Requirements Engineering for Business Process Support (REBPS 2003), Klagenfurt/Velden, Austria (2003)

[2] Attaran, M.: Exploring the relationship between information technology and business process reengineering. Information & Management 41, 585–596 (2003)

[3] Bleistein, S.: B-SCP: an integrated approach for validating alignment of organizational IT requirements with competitive business strategy. PhD Thesis, The University of New South Wales, Sidney, Australia (2006)

[4] Bubenko, J., Persson, A., Stirna, J.E.: User Guide (online) (2001), http://www.dsv.su.se/~js/ekd_user_guide.html

[5] CARE Technlogies, http://www.care-t.com

[6] Castro, J., Kolp, M., Mylopoulos, J.: Towards requirements-driven information systems engineering: the Tropos Project. Information Systems 27, 365–389 (2002)

[7] Constantine, L., Lockwood, L.: Software for Use: A Practical Guide to the Models and Methods of Usage- Centered Design. Addison Wesley, Reading (1999)

[8] Curtis, B., Kellner, M., Over, J.: Process Modelling. Communications of the ACM 35(9), 75–90 (1992)

[9] Dardenne, van Lamsweerde, A., Fickas, S.: Goal-directed Requirements Acquisition. Science of Computer Programming 20, 3–50 (1993)

[10] Daoudi, F., Nurcan, S.: A Benchmarking Framework for Methods to Design Flexible Business Process. Software Process Improvement and Practice 12, 51–63 (2007)

[11] Dobing, B., Parsoms, J.: Understanding the role of use cases in UML: a review and research agenda. Journal of Database Management 11(4), 28–36 (2000)

[12] Dumas, M., van der Aalst, W., ter Hofstede, A.: Process-Aware Information Systems. Wiley, Chichester (2005)

[13] Eriksson, H., Penker, M.: Business Modeling with UML: Business Patterns at Work. John Wiley and Sons, Chichester (2000)

[14] Estrada, H., et al.: An Empirical Evaluation if the i* Framework in a Model-Based Software Generation Environment. In: Dubois, E., Pohl, K. (eds.) CAiSE 2006. LNCS, vol. 4001, Springer, Heidelberg (2006)

[15] International Institute of Business Analysis. Business Analysis Body of Knowledge (online) (2006), http://www.iiba.com

[16] International Organization for Standardization. ISO 9001:2000, http://www.iso.ch

[17] Holtzblatt, K., Beyer, H.: Requirements gathering: the human factor. Communications of the ACM 38(5), 31–32 (1995)

[18] Lauesen, S.: Task Descriptions as Functional Requirements. IEEE Software 20(2), 58–65 (2003)

[19] Luftman, J., Raymond, R., Brier, T.: Enablers and Inhibitors of Business-IT Alignment. Communications of AIS 1 (1999)

[20] Lukaitis, S., Cybulski, J.: The Role of Stakeholder Understanding in Aligning IT with Business Objectives. In: REBNITA 2005, Paris, France (2005)

[21] Marshall, C.: Enterprise Modeling with UML. Addison-Wesley, Reading (2001)

[22] Nysetvold, A., Krogstie, J.: Assessing Business Process Modeling Languages Using a Generic Quality Framework. Advanced topics in database research 5, 79–93 (2006)
[23] OMG. Business Process Modeling Notation (BPMN) Specification (online) (2006), http://www.bpmn.org
[24] OO Method project, http://oomethod.dsic.upv.es
[25] Ould, M.: Business Processes: modelling and analysis for re-engineering and improvement. Wiley, Chichester (1995)
[26] Pastor, O., Molina, J.C.: Model-Driven Architecture in Practice: A Software Production Environment Based on Conceptual Modeling. Springer, Heidelberg (2007)
[27] Reich, B., Benbasat, I.: Factors That Influence the Social Dimension of Alignment Between Business and Information Technology. MIS Quarterly 24(1), 81–113 (2000)
[28] Rubens, J.: Business analysis and requirements engineering? The same, only different? Requirements Engineering 12, 121–123 (2007)
[29] Scheer, A.-W.: ARIS - Business Process Frameworks, 3rd edn. Springer, Berlin (1999)
[30] Siau, K.: The Psychology of Information Modeling. Advanced topics in database research 1, 116–118 (2002)
[31] Smith, H., Fingar, P.: Business Process Management: The Third Wave. Meghan-Kiffer Press (2002)
[32] Standish Group. Chaos, http://www.standishgorup.com
[33] Vessey, I., Coner, S.: Requirements Specification: Learning Object, Process, and Data Methodologies. Communications of the ACM 37(5), 102–113 (1994)
[34] WfMC. Workflow Management Coalition: Terminology & Glossary (1999)
[35] Wahl, T., Sindre, G.: An Analytical Evaluation of BPMN Using a Semiotic Quality Framework. Advanced topics in database research 5, 94–105 (2006)
[36] Wohed, P., et al.: On the suitability of BPMN for Business Process Modelling. In: Dustdar, S., Fiadeiro, J.L., Sheth, A.P. (eds.) BPM 2006. LNCS, vol. 4102, Springer, Heidelberg (2006)
[37] Yu, E.: Modeling Strategic Relationships for Process Reengineering. PhD Thesis, University of Toronto, Canada (1995)
[38] Zave, P., Jackson, M.: Four Dark Corners of Requirements Engineering. ACM Transactions on Software Engineering and Methodology 6(1), 1–30 (1997)

Weak Dependencies in Business Process Models

Marek Kowalkiewicz[1], Ruopeng Lu[1], Stefan Bäuerle[2], Marita Krümpelmann[2], and Sonia Lippe[1]

[1] SAP Research
{marek.kowalkiewicz,ruopeng.lu,sonia.lippe}@sap.com
http://www.sap.com/research
[2] SAP AG
{stefan.baeuerle,marita.kruempelmann}@sap.com
http://www.sap.com/

Abstract. Business process management systems (BPMS) have proven to be effective in high-volume, repetitive production processes. However, the rigid structure of process models used in BPMS hinders their use in less repetitive processes performed by information workers, where a high degree of flexibility is required. In this paper, an extension to the traditional process models is presented, where ad hoc, runtime changes to process instances are enabled. The extension, weak dependencies, is motivated by business cases presented in the paper. Additionally, formal description of weak dependencies as well as a proof-of-concept prototype are presented.

Keywords: ad hoc workflows, workflow modelling, flexible workflows.

1 Introduction

In commercial BPMS, graph-based workflow models have been prevalent as the mechanism to enforce control over process automation. The popularity of graph-based models is due to the intuitiveness and visual appeal of the graphical nature. Graph-based process modelling is an answer to a requirement of presenting business process models to various stakeholders in an as straightforward manner as possible.

Workflows have been delivered effectively in high-volume, repetitive processes, such as production line processes. A typical graphical workflow model (schema) is highly prescriptive, capturing (or trying to capture) every possible execution configuration. At runtime, multiple instances of the same workflow can be created according to the same workflow model. That approach allows for creating high volume of process instances from a limited number of workflow models. While viewed as a correct approach for high-volume repetitive processes, the rigidity of graphical workflow models has been considered as the main source of inflexibility in workflow modelling and execution. This is the case particularly for processes that demand greater flexibility, including ad hoc, knowledge-intensive processes, such as pathology diagnosis processes in healthcare domain, e-learning or others, where knowledge workers are involved.

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 177–188, 2008.

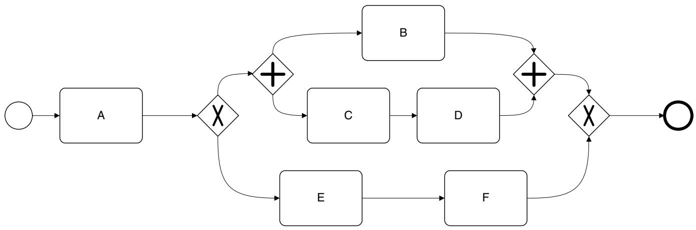


Fig. 1. A sample workflow model created using BPMN [7]

Many business scenarios demand flexibility in workflows, i.e. the ability to support the deployment of *partially specified* (underspecified) workflow models, where the complete model is *fully specified* (adapted) at runtime, according to instance-specific conditions [13]. The demand for flexibility is mainly due to two reasons. (1) Due to lack of complete a priori knowledge of the overall process during process design, e.g. a process may be expected to perform additional activities during execution. An incomplete workflow model may lead to frequent runtime exceptions resulting in changes to the deployed workflow model. (2) Business processes evolve over time due to the ever changing business requirements, such as the implementation of industrial best practices and government regulations (e.g. Sarbanes-Oxley Act).

There are BPMS that allow for modifying workflow instances during the execution. However, their use is not desirable from a business user point of view. The solutions are very general, and as such highly complex. With the current systems, it is not possible to specify the scope of a change (which parts of the process can be modified), the degree of change (what types of modifications are possible), or the agent of change (who is authorised to perform the modifications). As a result, the current solutions are a tool for skilled process modellers to respond to exceptions in processes, but do not provide any value to end users of processes, who might not even be able to understand the modelling notation.

Under this setting, we propose a solution where an underspecified workflow model can be defined for a particular business process at design time. At runtime, the workflow model for each workflow instance can be fully specified by adapting the general workflow model, according to specific needs and the definition of the process. In this paper, we present one facet of the solution, allowing inserting activities and groups of activities in specified areas of the model. We argue that such a basic extension is already addressing a large group of business cases and allows for modifying process instances without the need to understand the complex notation. As a result of our proposal, workflow models can be defined in a more flexible way, allowing different instances to determine their own unique processes.

Currently there is no such graph-based workflow modelling language that preserves the visual appeal of graphical models yet being expressive enough to support the above mentioned requirements. We believe that the approach presented in this paper paves the way for such extensions, showing that for specific cases of flexibility, the current approaches to workflow modelling can be reused. We also view this research as relevant, as the capability to change a business process immediately in response to changes of the requirements can solve a number of problems that information systems designers and users face.

In Section 2 of the paper, the related work is discussed. In Section 3, we provide a general description of weak dependencies, followed by a formal explanation and instructions how to implement weak dependencies in systems that do not support the construct. We finish the section by describing a proof-of-concept prototype that we have created to verify our claims. In Section 4, we summarise the contribution of the paper and discuss possible future work.

2 Related Work

BPMS are known by the lack of support for runtime flexibility, due to the typical single schema architecture. The requirements for providing flexibility in BPMN[1] arose from the need for change in workflows, which has been identified in the mid 90s [3,4]. The lack of support for runtime flexibility is due to several reasons.

Evolution of Process Requirements. Process goals and requirements are constantly changing, due to the dynamic business environment. Factors for change include changes in business goals, change in resource availability, incorporating best practices, and implementing compliance controls [13].

Tacitness of Process Knowledge. Process modellers lack comprehensive runtime knowledge at design time. The tacitness of process knowledge is due to the fact that knowledge constitutes the corporate skill base and is found in the experiences and practices of individual workers, who are domain experts in a particular aspect of the overall operations.

Limitation of Modelling Languages. The classic trade-off between expressibility of the modelling language and complexity of process model still applies. It is possible that business situations exist that cannot be specified in the process model due to the lack of expressive power. This would potentially lead to forced model derivations at runtime (such as system workarounds). On the other hand, the complexity in the latter case may compromise readability of the process model, and results in the added complexity in model verification.

Limitation of Process Execution. Typically, a process model is designed and deployed for each process type. It is common that process instances may deviate from modelled behaviours during execution, due to various runtime issues such as resource unavailability and evolving process requirements.

The issues of flexibility in BPMN involve issues such as flexibility, adaptability dynamism and evolution [12,15]. In general, the underlying requirement is the ability to provide support for change in BPMS. Changes in workflows can be classified into unanticipated and anticipated [15].

[1] In this paper, we have chosen BPMN as the notation for process models. However, the proposed solution can be demonstrated by other modelling languages.

Unanticipated changes refer to runtime modifications of process model, on either type or instance level in order to handle an unexpected situation. Typically, modification operations include adding or deleting an activity from the process model, or altering the execution pattern of one or more activities. The change operations are generally regulated by the correctness criteria for change [10,11], such that the resulting model is still structurally correct (e.g., deadlock free). ADEPT [10] is a research prototype for a flexible workflow execution environment that supports runtime structural modification.

Anticipated instance level change is regarded as the major strength of flexible workflows and has been receiving much attention in recent years. The introduction of flexible components into workflow models requires the ability of the business process to execute on the basis of a loosely, or partially specified model, where the full specification is made at runtime. An early attempt in this direction is the flexible sequence in MOBILE [5]. Since then, there have been many proposals offering various solutions, notably, Pocket of Flexibility [13], Worklets [2], Case Handling [1], and AgentWork [9]. The industry standard modelling language BPMN [7] also provides a construct called ad hoc sub-process to cater for such requirement. In the Worklets approach [2], a worklet is a discrete process fragment that is designed to handle a specific action (task) in a process. An extensible repository (repertoire) containing a number of different worklets is maintained for a worklet-enabled activity, such that at runtime a preferred worklet is contextually chosen to fulfil the activity goal. The selection of worklets is guided by a set of ripple-down-rules which associate a worklet with a series of instance-specific conditions. The Pocket of Flexibility approach [13] provides a placeholder activity which contains a set of unstructured inner activities. A fundamental feature is specification of build constraints which essentially control the modelling of the unstructured activities. Late modelling starts when the pocket is instantiated. Then a domain expert defines a corresponding process fragment using a restricted set of modelling elements. The inner activities can be modelled to execute in sequence or parallel, as long as the build constraints are not violated.

Weak dependency falls into the late modelling paradigm [15], which allows for selecting the implementation for a particular process step at runtime based on predefined rules or user decisions. Late modelling supports deploying a partially specified process schema, where parts of it have not been defined at design time, but are modelled at runtime for each process instance. The rationale of weak dependency is to support the cases when newly discovered workflow activities are required to dynamically modelled and inserted into running process instances, by an extension to the graphical modelling notation and the conceptualisation of the runtime process adaptation technique. A similar concept is the *empty activity* in BPEL [8]. An empty activity is used for inserting a no-operation instruction into a process, which can be considered as a placeholder for catching and suppressing exceptions. The distinction between the two is that weak dependency is designed to support flexible execution at instance level, while empty activity does not allow for flexible execution, as supported by the former.

3 Weak Dependency

In this section, we introduce the concept of weak dependency, improving the flexibility of business process execution. We present the general idea of the solution along with a formal description of the method, and we show a prototype implementation of the solution. Before discussing weak dependencies, we introduce basic workflow concepts of workflows in Section 3.1, on which we build upon later in the text.

3.1 Workflows and Process Models

We consider a workflow graph as a directed-acyclic graph (DAG), consisting of a set of nodes and a set of directed edges, where a node represents either an activity in the workflow, an event, or a gateway, modelling a control flow construct. An activity is either an automated or manual activity. An event is either an initial event (start event) or a final event (end event). A gateway, modelling a control flow construct can be of type *sequence*, *and*, or *exclusive or*. The *and* and *exclusive or* gateways can act as either a split gateway or a join gateway. For the sake of simplicity, we do not consider other types of gateways in this paper. Below, we provide a formal definition of a workflow [16].

Definition 1. *Let C be a set of control flow constructs. $P = (N, E, type)$ is a* process model *if it consists of a set of N nodes, and a set of E edges.*

- *$N = N_A \cup N_E \cup N_G$, where N_A is a set of activity models, N_E is a set of event models, and N_G is a set of gateway models. These sets are mutually disjoint.*
- *E is a set of directed edges between nodes, such that $E \subseteq N \times N$, representing control flow.*
- *$type : N_G \mapsto C$ assigns to each gateway model a control flow construct.*

An edge $e \in E$ connects two nodes (activities, events or gateways), indicating immediate precedence relationship, which imposes two types of requirements:

1. the node at the source of the edge is the precondition of the node at the destination of the edge (it will not be enabled, unless the node located at the source has terminated),
2. the latter node occurs (if it is an event node) or is enabled (if it is an activity or a gateway) immediately after the prior (there is no delay, no other nodes can be present in between the two).

That behaviour of the edge is referred to as *strong dependency* in this document. In a typical workflow graph, each directed edge represents strong dependency between the pair of nodes it connects (cf. Fig. 2). Such workflow model has a descriptive nature in that the completeness of the model is assumed at design time, where model modification after deployment is unanticipated. As a result, workflow models with only strong dependencies are not suitable to express flexible workflows.

Fig. 2. Strong dependency between activities A and B

3.2 Weak Dependencies at a Glance

In dynamic business environments, different workflow instances may require run-time adaptation to the workflow model, specifically to insert new nodes in the process which were not anticipated at design time. These modifications require changing the structure of the workflow (instance) graph. Consider the situation when one or more additional nodes have to be added between activities A and B in Fig. 2. It is desirable to indicate that there is an "editable" part in the process model between activity models A and B, while keeping A and B still connected, expressing their ordering constraints (B can be enabled only after A has terminated, however not requiring that to happen immediately afterwards). This scenario is not expressible in general workflow models (Fig. 3).

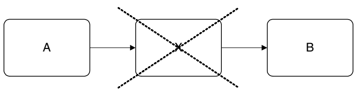


Fig. 3. Strong dependency cannot express requirement for dynamic activity insertion. No new activities can be inserted at runtime.

In order to overcome that clear shortcoming of current business process modelling notations, we introduce a new concept, *weak dependency*. Weak dependency is introduced as an enhancement to the graphical workflow model to express a weaker notion of control. A weak dependency between two activities in a workflow, e.g. between A and B, indicates that A is the predecessor of B (termination of A is the pre-condition of enabling B), while additional nodes can be inserted between them. To include weak dependencies in a definition of a process model, we propose the following change to the definition.

Definition 2. *A* process model with weak dependencies *is a* process model *in which there are two types of edges used: strong dependencies and weak dependencies.* $E = E_S \cup E_W$ *is a set of directed edges between nodes, where* E_S *is a set of strong dependencies, and* E_W *is a set of weak dependencies.* $E \subseteq N \times N$ *represents control flow.* $E_S \cap E_W = \emptyset$, $E_W \neq \emptyset$

For the purpose of this paper, notation of weak dependency is a dashed line with arrowhead (cf. Fig. 4), in contrast to the standard solid line with arrowhead used for strong dependencies in BPMN. Fig. 4 shows a fragment of a workflow model with weak dependency between activities A and B. Fig. 5 is a possible adaptation of the instance during runtime, where an additional activity, X, is placed between activities A and B.

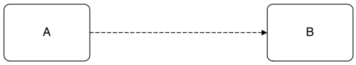


Fig. 4. Weak dependency between activities A and B

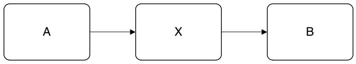


Fig. 5. Possible runtime adaptation—activity X is inserted between A and B

Weak dependency expresses the connection between two nodes, denoting only the interdependency between them, while immediate placement is not mandatory. Apart from activities, other types of nodes can also be introduced in place of a weak dependency when adapting the process. Fig. 6 shows an example workflow model with weak dependencies, which indicates three "editable" parts of the model—after activity A, before activity B, and between activities E and F. This flexible workflow model can be deployed in a BPMS which supports instance adaptation. When a workflow instance is created at runtime, an end-user (domain expert)[2] is involved to adapt the workflow model according to instance specific conditions, such that the flexible parts of the model are specified in full. This type of runtime adaptation is usually called process adaptation, and the adapted workflow model is referred to as a full specification of the flexible model. Fig. 9 shows a possible scenario of adaptation for the model from Fig. 6.

Implementation of weak dependencies in currently existing BPMS is achieved by replacing each of them with a *management tasks* and a pair of strong dependencies (Fig. 7). The management task is responsible for initiating user interaction, during which the relevant part of the workflow, represented by the weak dependency, is adapted. The adaptation phase may be skipped (weak dependency is changed to a strong dependency and the subsequent activity is enabled) or new activities and combinations of them may be added. The management task is active as long as the user wants to adapt the weak dependency, or until a predefined time has elapsed.

In Fig. 8 we show a simplified state diagram, representing the execution of an instance of a process with a weak dependency (cf. Fig. 7). Each activity is represented as a line segment, where the beginning of each segment represents the *enabled* state, and the end of it represents the *terminated* state. First, activity A is enabled, and after it terminates, the *Management task* is enabled. As a part of the task's execution, a user selects two activities (C and D), and after they have terminated, the user decides to execute additional ones (E, F, G). After the last activity has terminated, and the user does not want to add any more activities as part of the weak dependency (or after a predefined time passes, disabling adding any new activities), the *Management task* terminates, and activity B is enabled. Note that the simplified event diagram is not expressive enough to show how the

[2] Resource allocations and relevant strategies are beyound the scope of this paper.

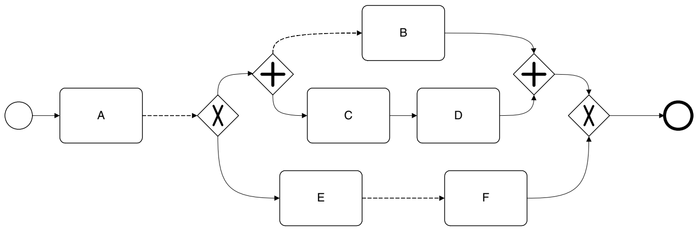


Fig. 6. Example process model with weak dependencies

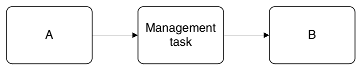


Fig. 7. Implementation of weak dependencies as a management task (cf. Fig 4)

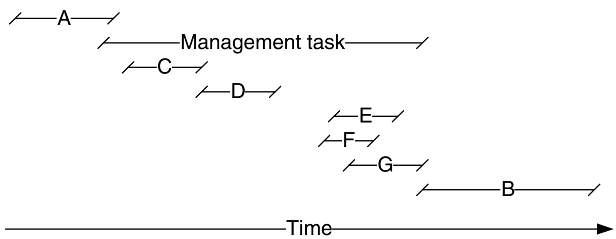


Fig. 8. Simplified event diagram of a weak dependency implemented as a management task with several activities inserted (cf. Fig 7)

additional activities were modelled by the user, for instance, whether activities C and D were modelled as a sequence or as running in parallel.

Further on, we show sample adaptations of a workflow, represented in Fig. 6. In the adaptation shown in Fig. 9, a sub-process G and H is added before B where G and H are executed in sequence; another sub-process containing two parallel activities I and J is added between E and F.

3.3 Prototype

To verify our approach, we have built a proof-of-concept prototype, running on a BPMS, that does not natively support weak dependencies. We have modelled weak dependencies as management tasks, and created a micro-modelling environment that can be accessed by non-expert users to adapt weak dependencies during the execution of workflows.

Consider an example, where a process for preparing and advertising new rental properties is executed. We show the sample process in Fig. 10. The process model shown is already a refined version of the original process, where one of the strong dependencies has been replaced with a weak dependency. A reason for that is that *Clean the property* may reveal unexpected conditions of the

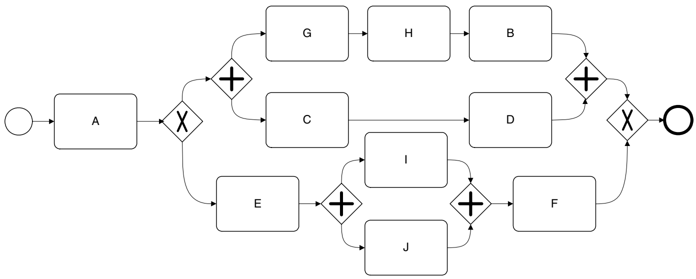


Fig. 9. Possible adaptation of the flexible model in Fig. 6

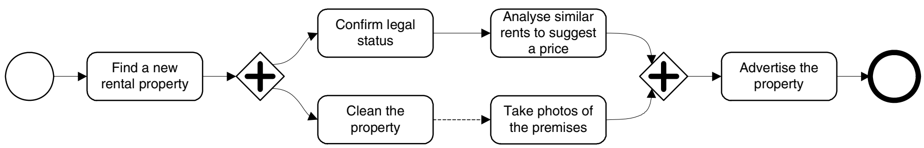


Fig. 10. A sample process, where weak dependencies may be required

property (pested, leaking pipes etc.) and it is unreasonable to try to model reactions to each possible situation. Additionally, performing additional tasks (for example fixing pipes) may cause situations that require further, unexpected, actions (for instance second cleaning). In such situation, it is not possible to model that effectively, as any approaches to model all possible situations will lead to explosion of the model. Even if the list of available services is fairly low, it is cumbersome to model situations where a number of activities are executed in parallel and decisions on which additional activities are to be performed, if any, are made on the fly. In such cases, weak dependencies are the proper approach.

In Fig. 11, we show a user interaction which occurred due to existence of a weak dependency, as in Fig. 10 connecting two activities: *Clean property* and *Take photos* in a process model. The user interaction screen shows immediate context of the modelling (adaptation) decisions, so that the end users know that the activities that they model will be executed in between the two activities. The user interaction screen allows for a simple inclusion of additional activities as well as provides two basic constructs—parallel and sequential placement of activities. Should the user decide to skip the adaptation, the *Finalize* button can be pressed without introducing any new tasks. In such case, the weak dependency will effectively be replaced with strong dependency, management task terminated, and *Take photos* enabled.

In this use case, the application of weak dependency is clear. The *Clean property* activity may reveal problems with the property that need to be resolved, some of them preventing making proper photos of the property. Therefore, the user is presented with a simple modelling environment, where activities from a

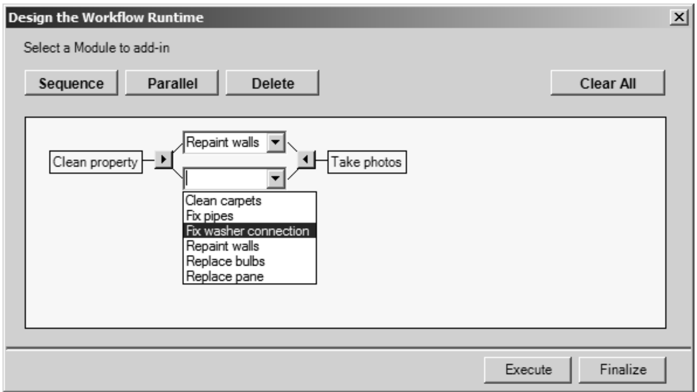


Fig. 11. Runtime adaptation. The management task has been enabled, and a user can specify how the weak dependency is adapted.

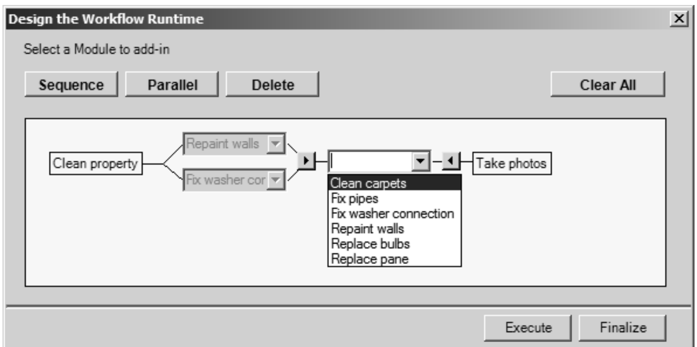


Fig. 12. Runtime adaptation. After executing the first phase of the adaptation, the user can adapt further or finalise the adaptation.

repository can be chosen, modelled, and enacted. This process can be repeated as many times as needed, since performing new activities may reveal further problems. When the user decides that there are no additional activities to be performed, the *Finalize* button enables the next activity. Otherwise, the *Execute* button enacts already modelled activities, and once they have finished, the end user is presented with further possibilities to add new activities (Fig. 12).

Since the target users of the simplified modelling environment are business users who need to make quick and simple decisions, we have decided to keep the modelling environment as easy to use as possible. We expect that in almost all cases, the adaptation will involve a very small number of new activities. Additionally, due to the nature of the adaptation (ability to go back to the adaptation screen after a few activities have executed), there is no need to introduce more complex constructs, such as decision gateways etc. The notation is straightforward and throughout the tests we have not come across users who could not understand it or understood it improperly[3].

In Fig. 12, we show the second phase of the adaptation, where the user has not finalised the adaptation phase in the first step. The activities that had already executed have been greyed out, to show the extended contextual information of the weak dependency, and new activities can be inserted only to the right of them. Depending on user choice, after the last modelled activity has terminated, the management task window will be brought up again, or the task will be terminated, and *Take photos* enabled. In this case, repainting walls and fixing water connection has resulted with stains on carpets, therefore an additional activity, *Clean carpets* has to be performed.

It is important to mention that in our approach we are using a repository of activities. The repository can be extended or modified at any time, and—as the drop down lists are populated directly from the repository—all changes will instantly be visible to end users. Therefore, no need to modify underlying process models occurs when new activities are introduced. In our study we were not considering how a large number of activities to choose from would affect the user experience. However, one may assume that proper filtering or ranking techniques should be applied in such cases.

4 Summary and Outlook

Flexibility has been one of the major challenges for current BPMS uptake. We consider an essential requirement for flexibility as the ability to execute on the basis of a loosely specified process model, where the full specification is made at runtime, and may be unique to each instance. In this paper we have introduced the concept of weak dependency, in respond to such a requirement. We have further provided formal specification of related concepts and shown how weak dependencies can be implemented in BPMS that do not support such construct. Finally, a proof-of-concept application has also been presented.

Weak dependencies for graphical workflow models can be applied in most graph-based process modelling languages supported by commercial BPMS or Process Aware Information Systems (PAIS). The notation presents a weaker requirement of sequence execution, and visualizes the editable region(s) of a workflow model for controlled runtime model modification. Examples of applications can be found in a wide range of business and workflow domains. To name a few, in knowledge intensive workflows such as patient treatment process in healthcare domain and network diagnosis process in telco domain, where the balance is kept between flexibility for individualizing process instances, and maintaining control of the overall process.

There is a number of further research topics that arise from this work. One interesting aspect to explore is the data dependencies with regard to weak dependency. The flexibility for modification on structural (control flow) level inevitably affect data binding in the process.

[3] However, formal study of understandability of the notation could be performed.

References

1. van der Aalst, W.M.P., Weske, M.: Case handling: a new paradigm for business process support. Data & Knowledge Engineering 53(2), 129–162 (2005)
2. Adams, M., ter Hofstede, A.H.M., Edmond, D., van der Aalst, W.M.P.: Implementing Dynamic Flexibility in Workflows using Worklets. BPM Center Report, Queensland University of Technology (2006)
3. Casati, F., Grefen, P., Pernici, B., Pozzi, G., Sánchez, G.: A Specification Language for the WIDE Workflow Model. Technical report, University of Twente (1996)
4. Ellis, C.A., Keddara, K., Rozenberg, G.: Dynamic change within workflow systems. In: Conference on Organizational Computing Systems, COOCS 1995, pp. 10–21 (1995)
5. Jablonski, S.: MOBILE: A Modular Workflow Model and Architecture. In: Proceedings of Int'l Working Conference on Dynamic Modelling and Information Systems, Nordwijkerhout (1994)
6. Jablonski, S., Bussler, C.: Workflow management - Modelling concepts, architecture and implementation. International Thomson Computer Press (1996)
7. Object Management Group (OMG): Business Process Modeling Notation (BPMN) Specification 1.0 (2006)
8. OASIS: Business Process Execution Language for Web Services (BPEL4WS) Version 1.1. Specification Organization for the Advancement of Structured Information Standards (OASIS) (2006)
9. Müller, R., Greiner, U., Rahm, E.: AGENT WORK: a workflow system supporting rule-based workflow adaptation. Data Knowl. Eng., Elsevier Science Publishers (51), 223–256 (2004)
10. Reichert, M., Dadam, P.: ADEPTflex—Supporting Dynamic Changes of Workflows Without Losing Control. Journal of Intelligent Information Systems 10(2), 93–129 (1998)
11. Rinderle, S., Reichert, M., Dadam, P.: Correctness criteria for dynamic changes in workflow systems - a Survey. Data Knowl. Eng. 50(1), 9–34 (2004)
12. Sadiq, S., Marjanovic, O., Orlowska, M.E.: Managing Change and Time in Dynamic Workflow Processes. Int. J. Cooperative Inf. Syst. 9(1-2), 93–116 (2000)
13. Sadiq, S., Orlowska, M.E., Sadiq, W.: Specification and validation of process constraints for flexible workflows. Information Systems 30 (2005)
14. Sadiq, S., Governatori, G., Namiri, K.: Modeling Control Objectives for Business Process Compliance. In: Alonso, G., Dadam, P., Rosemann, M. (eds.) BPM 2007. LNCS, vol. 4714, pp. 149–164. Springer, Heidelberg (2007)
15. Weber, B., Rinderle, S., Reichert, M.: Change Patterns and Change Support Features in Process-Aware Information Systems. In: Krogstie, J., Opdahl, A., Sindre, G. (eds.) CAiSE 2007 and WES 2007. LNCS, vol. 4495, pp. 574–588. Springer, Heidelberg (2007)
16. Weske, M.: Business Process Management: Concepts, Languages, Architectures. Springer, Berlin/Heidelberg (2007)

Advanced Querying and Reasoning on Business Process Models

Ivan Markovic

SAP Research
Vincenz-Priessnitz-Strasse 1, 76131 Karlsruhe, Germany
ivan.markovic@sap.com

Abstract. In order to respond quickly to changing market requirements, a business organisation needs to increase the level of agility in all phases of the business process engineering chain. Business process (BP) modelling is the first and most important phase in this chain. Designing a new and redesigning an existing process model is a highly complex, time consuming and error prone task. The research question that this work investigates is how to facilitate the design of new and redesign of existing process models by utilizing Semantic Web technologies. We present an approach for querying and reasoning on business process models which i) supports decision making, ii) facilitates reuse of modelling artefacts and iii) helps ensuring compliance of models to relevant regulations.

1 Introduction

In the modern world, businesses constantly strive to reinvent and differentiate themselves under continuous pressures of regulatory and technological change, as well as the increasing time to market requirements. One of the main obstacles for these changes to be agile is the lack of support when incorporating new business requirements into existing information systems as priorities and perspectives change.

Business process (BP) modelling is the first and most important phase in the business process engineering chain. BP models are created by business analysts with an objective to capture business requirements, enable a better understanding of business processes, facilitate communication between business analysts and IT experts, identify process improvement options and serve as a basis for derivation of executable business processes. Designing a new process model is a highly complex, time consuming and error prone task. This is because BP modelling requires usage of various sources of information, models are frequently redesigned to adapt to changes, and BP models are often shared by several departments within a company or even between different companies.

In order to simplify BP modelling, models need to be highly reusable, favoring process flexibility and minimizing designs made from scratch. Reusing implies the need for querying a process repository in order to find suitable previous work that can be the base for a new design. This can be done only by using an expressive and machine-readable description of relevant aspects of a BP model

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 189–200, 2008.

that will help to retrieve the most relevant parts of a previous work (model). Following this idea we: i) create a formal model for describing business processes which integrates different workflow perspectives, and ii) devise an approach for expressive querying and reasoning on business process models.

The remainder of the paper is organized as follows. In Section 2 we provide a requirements analysis for creating a framework for querying and reasoning on process models. Section 3 reports on our approach for representing, querying and reasoning on business processes. Section 4 discusses related work. We conclude and give an outlook on future research in Section 5.

2 Requirements Analysis

In this section we describe the usage scenarios that a framework designed for querying and reasoning on business process models should support and derive a list of requirements that the framework should fulfill. These scenarios and requirements were partly derived from the interviews conducted with business experts concerned with process modelling.

2.1 Usage Scenarios

In Figure 1 we depict the usage scenarios considered in this work:

– *Scenario 1: Decision making support*
  The key challenge in decision making is having access to all relevant information which is to be assessed in a particular situation. Such information is scattered in organization processes and has to be manually collected from diverse sources for each individual case. To facilitate this task, we enable the business expert to quickly and expressively query the process artefact

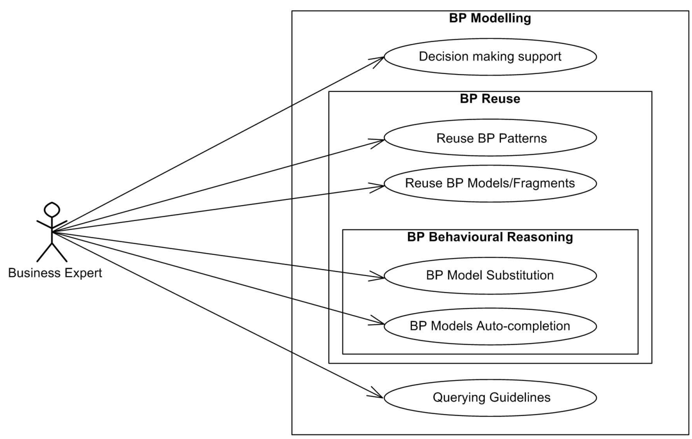


Fig. 1. Usage scenarios

repository (process patterns, models, fragments and modeling guidelines) of an organization (cf. Figure 1, top). Some example queries for this scenario include: "*Give me all processes in the fulfillment area*", "*Which processes use system x?*", "*What resources are needed for running process y?*", "*List all processes with conflicting goals.*", "*Which processes support the achievement of business goal z?*" [1].

– *Scenario 2: Reuse of process artefacts*
  This scenario describes how can a business expert query the process artefact repository for reuse of process patterns, models and fragments in process design (cf. Figure 1, center). Since process modelling is a complex activity, reuse of existing models and model components makes sense in all stages of modelling. For instance, when designing a new process the business expert can first query for existing business process patterns, generic high level process designs emphasizing business goals [2], in search for the best modelling practices in the given domain. An example query for business patterns can be: "*Give me all business patterns related to Fulfillment Business Function where Business Goals involved are profileObtained and serviceActivated*". The business expert can also query in the same way for existing models or process fragments - self-contained, coherent building blocks of a process model with a clear business meaning. In case that there are existing process models or fragments that are similar to the desired end design, the business expert can use them in his design in order to achieve a higher degree of reuse, compared to reuse of patterns. Moreover, if the user wants to substitute an existing process fragment based on redesign goals or auto-complete an underspecified model, he can make graphical queries by selecting the desired process part for substitution or auto-completion in the modelling tool. For this purpose we use properties of bisimulation theory for the π-calculus [3] (cf. Figure 1, center).
– *Scenario 3: Querying modelling guidelines*
  This scenario covers querying for business guidelines - concrete policies defined according to the company strategy, which apply orthogonally to all processes of an organization (cf. Figure 1, bottom). Queries involved in this scenario retrieve all modelling guidelines (both mandatory and conditional) which match context annotations of the model being checked. Process context is composed out of the following perspectives: Business Function, Business Domain, and Business Goal, which are modeled in the appropriate process ontologies. Querying for modelling guidelines based on process context information reduces the manual effort of creating an inventory of such guidelines for any given model. For checking which guidelines are relevant in a digital content provisioning process, the example query can be: "*Give me all modelling guidelines for Digital Asset Management Business Function where clients are minors and Business Goal associated belongs to Fulfillment*". For more details on different types of queries that business expert can perform in process modelling, we refer the reader to [2].

2.2 Requirements

- *Req. 1: Rich process description* The process model needs to be formally described to enable automatic matchmaking of user requests (goals) against process descriptions. In order to support the user to expressively search the process repository, we need a rich process description. We distinguish two main aspects of a process description: dynamic and static aspect. Within the dynamic aspect we want to capture the behavior of the process, i.e. process control flow. This will give the user a possibility to impose behavioral constraints on the process artefacts he wants to retrieve. Within the static aspect of the process description we want to describe other workflow perspectives, e.g. organizational and informational. We want to describe processes in terms of their input/output data, business function, business domain, organizational roles which perform certain process parts, etc. In this way, the user will be able to specify this type of information in his request and use it to express additional constraints in his query.
- *Req. 2: Intuitive user request specification* The user must be provided with a user-friendly query interface for specifying his requests. The user must be able to query for processes both on the static and dynamic aspects of their description.
- *Req. 3: Query language* There needs to be a query language with expressive power that is sufficient to formally describe the user requests. Note that the user request can be in the form of a query template, but it can also be a part of a process for auto-completion or a process fragment for substitution.
- *Req. 4: Querying mechanism* There needs to be a mechanism that will perform expressive matchmaking of user requests against process descriptions. The algorithm should take a user request as an input and return a ranked list of results that match the request.
- *Req. 5: Flexibility* The framework must provide support for relaxation and refinement of user queries. In the case that we get too less results matching the query, we can relax the query, i.e. incrementally abstract elements of the query, e.g. by using subsumption hierarchy. Similarly, if we get too many results, the user needs to be provided with a possibility to refine his request.
- *Req. 6: Ranking* To increase the usability of matching results, the list of results should be ordered w.r.t. the level of match. Similarity measures need to be defined for process models to support the ranking of results based on multiple criteria.
- *Req. 7: Computational efficiency* A lesser requirement, yet the underlying querying mechanism must be computationally tractable and efficient to provide the required design-time user support.

3 The Approach

In this section, we present how we addressed the requirements listed in Section 2.2. We first discuss the formal model for business processes which is the basis of our approach. Second, we present the query language which has been defined for coupling static and dynamic process characteristics. Further, we shortly describe the mechanism for processing queries. In the last subsection we explain how we have introduced flexibility in querying for providing better query results.

3.1 Rich Process Description

In order to enable expressive querying and reasoning on BP models, there is a need for a comprehensive formal process model description capturing all relevant dimensions (perspectives) of a process. Following [4] and [5], we consider functional, behavioural, organizational and informational perspective relevant to adequately organize information about a process. Based on these requirements, in our previous work [6] we have proposed a formal model for describing business processes which integrates all aforementioned perspectives, as illustrated in Figure 2. In the following we are briefly describing these perspectives.

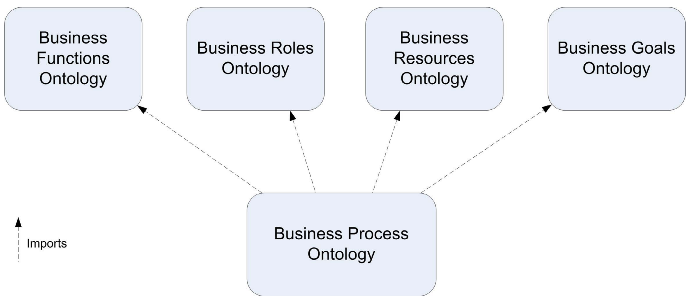


Fig. 2. Ontology framework: Perspectives for describing a business process

For describing the behavioural (dynamic) perspective of a process model we utilize process algebra, the π-calculus. There are several reasons for selecting the π-calculus as theory for describing the distributed and dynamic nature of modern Business Process Management (BPM) systems. First, the theory supports message-based interaction, an important requirement for supporting both intra- and inter-organizational processes [7]. Second, it supports the trend of shifting from central to distributed BPM systems and thus enables reasoning over distributed processes. Third, it supports the BPM shift towards dynamic, open environments with constantly changing number of interaction partners, such as the internet, using link passing mobility property. In addition, recent research has shown that π-calculus is able to easily express the formal semantics of all documented workflow patterns [8] as well as represent powerful service choreographies [9]. By using the π-calculus for representing the process behaviour, we are also able to integrate existing tools and techniques for verification and simulation of processes [10], [11] in our framework. The dynamic perspective of a process model stands for process control- and dataflow, and we model it using the ontologized π-calculus, denoted by Business Process Ontology in Figure 2. For more details on this ontology, we refer the reader to [6].

For representing the functional, organizational and informational perspective we have proposed a set of ontologies, imported by Business Process Ontology (BPO), as shown in Figure 2. Business Functions Ontology provides a structural breakdown of the organization's business functions. Concepts from this ontology classify process models by their functionality, independent of the business domain. Business Roles Ontology includes concepts representing roles in the organization, e.g. Manager, Engineer, Clerk, Secretary, etc. Business Resources Ontology describes the resources (documents, systems, machines) which are required to operate the activities in processes. Business Goals Ontology models a hierarchy of organization business goals (milestones, objectives) according to which the processes in the organization are designed. Business goals are modeled in such a way that they conflict if they can not be satisfied simultaneously. Moreover, goals can influence positively or negatively other goals [12]. Note that we refer to these perspectives as static view of a process in the rest of the text.

3.2 User-Friendly Query Interface

The intended users of our framework are business experts. Therefore, they need to be provided with an intuitive and user-friendly query interface to be able to specify their queries in an easy way, not much different from using the applications they are used to. The complexity of ontologies and reasoning needs to be hidden from the user. In the following, we describe how we designed the query interface based on this requirement.

The query input dialog for performing the queries on the static view of a process (static queries) is presented in Figure 3. The user can navigate through tabs for selecting business annotations (business goals, functions, roles and resources) of the processes he wants to retrieve. Note that these characterictics correspond to the perspectives depicted in Figure 2. In the right box, the ontology navigator provides available ontology concepts which the user can browse.

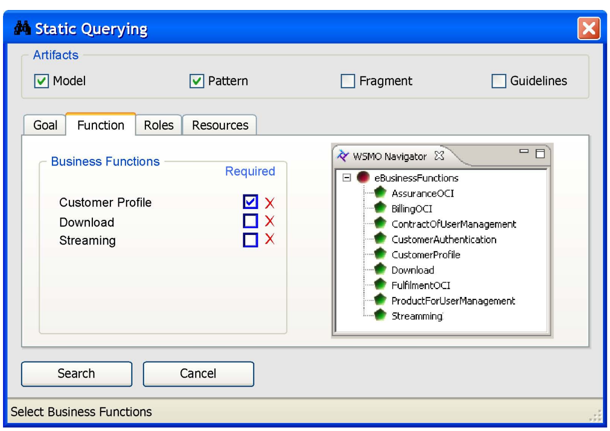


Fig. 3. Static querying interface

Desired business annotations can be dragged from the ontology navigator to the left box (Business Functions panel) and marked as required or optional for querying. This is important for achieving flexibility in querying, since the concepts marked as optional can be omitted when constructing the query for retrieving more results (cf. Section 3.4). The user can also specify the type of artefacts he is looking for (model, pattern, fragment, guideline) in the upper panel.

In contrast to performing static queries, querying for desired behaviour does not need a new dialog. Querying is performed by selecting a process part directly in the modelling tool, as shown in Figure 4 (shaded area). We consider this way of graphical querying to be intuitive to the user. The behavioural description of the selected part is obtained automatically using the Parser/Serializer component and used as an input for behavioural reasoning (substitution, auto-completion, deadlock/liveness verification).

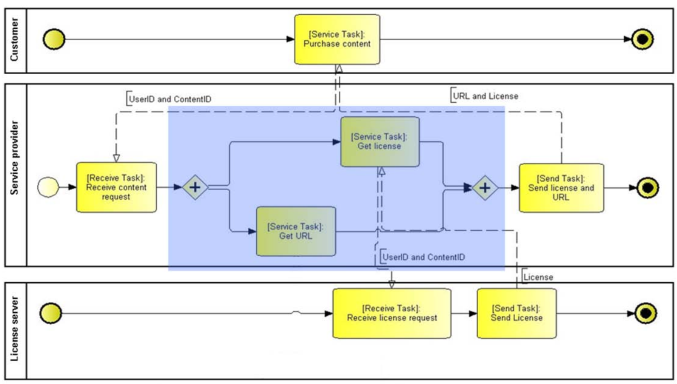


Fig. 4. Dynamic querying interface

3.3 Querying

In this section, we discuss how the querying process operates inside our framework. We first discuss the formal query language which has been defined for coupling static and dynamic process characteristics. Further, we shortly describe the mechanism for processing queries.

Formal query specification. The query must support the description of static and dynamic behaviour. WSML Logical Expression (LE) is used for specifying queries on static properties of a process. The syntax of WSML-LE is based on Frame Logic [13]. The query should reference instances of business annotations, which are materialized by instances of the relations described in [6]. An example is given in the following:

$$bpo\#hasBusinessFunction(?x, Annot1) \wedge bpo\#hasBusinessRole(?x, Annot2) \quad (1)$$

The dynamic behaviour of a wanted process is described as a process definition, i.e., it uses the ontology framework given in Figure 2 for describing a process, its connections, and the annotations of tasks as relation instances. Defining that the behavioural query has the same structure as the process definitions avoids the mismatch problem that would appear from having a different query language for behavioural querying.

Therefore, here we use the ontologized π-calculus to describe the user request (process query). This query is checked against processes stored in the repository using congruence and bisimulation properties.

The query specification must be defined to use the same language used by processes description. This enables the reuse of the ontological process model for representing user queries. For meeting our requirements, the query specification should be in a format of a template with placeholders. In addition, it should encapsulate both static and dynamic attributes in a unified language.

The use of a BPO ontology instance is the solution to this: the query template corresponds to a pre-defined ontology structure with namespace definition and element descriptions. Table 1 shows the template contents:

Table 1. Query definition using BPO ontology instance

Namespace	http://www.ip-super.org/ontologies/BPO/extension/query
dc:type	Either "substitution" or "auto-completion"
Axiom	ID:http://www.ip-super.org/ontologies/BPO/extension/query#static definedBy: bpo#hasBusinessResource(?x, *PLACEHOLDER*) 'OR' bpo#hasBusinessFunction(?x, *PLACEHOLDER*) 'OR' bpo#hasBusinessGoal(?x, *PLACEHOLDER*) 'OR' bpo#hasBusinessRole(?x, *PLACEHOLDER*)
Process	ID:http://www.ip-super.org/ontologies/BPO/extension/query#process A BPO process model description.

The ontology contains an axiom (containing the static query), and a process definition (containing the process behaviour). The non-functional property dc:type indicates the type of behavioural query being performed.

The ontology submitted as query contains all necessary information for specifying the query request. The approach enables performing ontological reasoning on the static part and π-reasoning on the dynamic part of process description. This approach is also scalable, since adding new concepts or adding new information to the query will not imply changes to the established query definition.

Query mechanism. In order to reduce the level of complexity, we have divided this task in two subtasks - static and dynamic (behavioural) querying. The querying mechanism operates on the ontology framework presented in Figure 2.

First subtask investigates simple (static) querying, where the user can specify constraints related to the static view of a process. Here we use WSML logical expressions as a query language and ontological reasoning for query answering.

The second subtask investigates graphical (behavioural) querying, where the user can specify requirements on dynamic perspective of a process description. This corresponds to auto-completion and substitution scenarios where algorithms from bisimulation theory are used for comparing the processes. For more details on the framework querying mechanism, we refer the reader to [2].

3.4 Achieving Flexibility in Querying

Depending on the user query, there can be too many or too few results coming from the repository. The user should be able to i) specify further constraints in the query, choosing more refined goals, thus retrieving more precise and shorter list of results or ii) eliminate some constraints from the query and look for more abstract goals or more undefined flow structure of the process, thus retrieving more results.

(i) The refinement is done mainly by the navigation inside sub-concepts, skipping instances connected to super-concepts. Since a concept can have many sub-concepts, the interaction with the user may be necessary to choose which path to follow. In Figure 5, an example of refinement is represented by a search for the goal "*Creation Done*", which has at the moment no instances directly connected, however, its sub-concepts "*User Profile Created*" and "*Catalog Entry Created*" do have, and processes associated with these instances are results of a refined query.

(ii) If there are too few results, the framework can automatically search for instances connected to the parent concepts of the requested one in order to relax the query. For example in Figure 5, a query for processes related to the goal "*Client Known*" has just one instance as a result. If the framework decides to look for instances connected to "*User Known*", two instances match the query.

In general, when going deeper in the tree structure of goals one would retrieve fewer results (refinement), while when going upward in the tree structure the

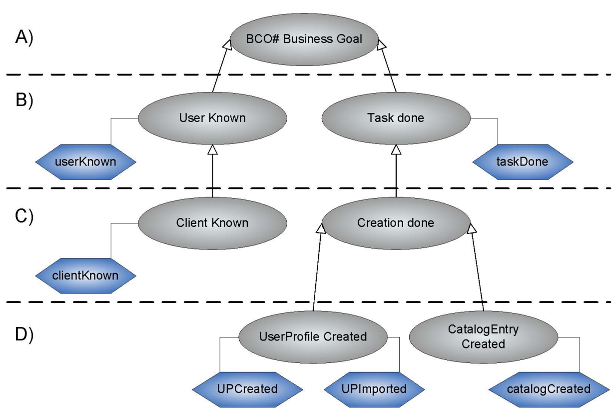


Fig. 5. Business goals hierarchy

relaxation occurs. A second way for relaxing a query is to "skip" target concepts in the query, e.g. using the "OR" statements. The user can specify which concepts are required and which are optional using the interface in Figure 3. Consider an example when the user wants models achieving goals A, B and C, where A and B are required to be parts of the model and C is optional. The framework can try to find models achieving all goals with the statement:

$$Q = \{x | x \in M \wedge ant\,(x, A) \wedge ant\,(x, B) \wedge ant\,(x, C)\} \tag{2}$$

Where Q represents the resulting set of process models, M denotes the set of all process models and relation ant denotes that a model x is annotated using the ontology concept A.

In case of no result, the framework decides to relax the query, executing the statement:

$$Q = \{x | x \in M \wedge ant\,(x, A) \wedge ant\,(x, B)\} \tag{3}$$

In the ideal case, the framework would submit the following statement, preferring to answer the user queries where all three goals are achieved:

$$Q = \{x | x \in M \wedge (ant\,(x, A) \wedge ant\,(x, B) \wedge ant\,(x, C)) \vee (ant\,(x, A) \wedge ant\,(x, B))\} \tag{4}$$

Here we see a clear need for ranking: results that fulfil all goals are matching perfectly the user query, while results obtained for the relaxed query are fulfilling just a part of the user's need. At present, the ranking of query results is not supported by our approach. That will be a part of the future work.

4 Related Work

The significance of querying business processes has been acknowledged by BPMI[1] who launched a Business Process Query Language (BPQL) initiative [14]. However, no standard specification has been published yet.

In [15], the authors present a query language for querying business processes, BP-QL. The query language is designed based on the BPEL[2] standard and thus focuses on querying executable processes. Our work focuses on the reuse of higher level business process knowledge, i.e. BP artefacts. In addition, our query specification language is more expressive in that apart from constraints on data and control flow, the user can specify additional properties of the models he wants to retrieve.

The approach presented in [16] discusses a process component model for process knowledge reuse. Here, the process component model is characterized only using static information (domain, function, performance, lifecycle). We defined a more refined notion of BP artefacts which can be reused in different stages of the presented modelling lifecycle. In addition, with this approach the user is not able to specify behavioural queries which is a more intuitive way of specifying his requests.

[1] Business Process Management Initiative, http://www.bpmi.org/

[2] Business Process Execution Language.

We have not seen other approaches addressing the problem of querying of business process models using a rich formal model for business processes.

5 Conclusion

In this work we have presented an approach which enables expressive querying and reasoning on business process models. The approach enables the business expert to have a quick and easy access to the library of process artefacts. We have illustrated three key usage scenarios showing the benefits of using our querying approach. Furthermore, we have developed a first prototype of the querying framework, based on the Maestro BPM tool. Currently we are performing a use case study and the evaluation of the results is the very next task.

As our next step, we plan to embed Mobility Workbench (MWB) [17] as a subsystem in our framework implementing bisimulation for the π-calculus. This will allow us to investigate more advanced notions of process equivalence based on the theory of bisimulation. In the long term, we aim to define similarity measures for process models with the purpose of quantifying the level of similarity between two models, which can be used in ranking of query results.

Acknowledgments

This work has been partially supported by the European Union within the FP6 IST project SUPER[3] (FP6-026850).

References

1. Hepp, M., Leymann, F., Domingue, J., Wahler, A., Fensel, D.: Semantic business process management: A vision towards using semantic web services for business process management. In: Lau, F.C.M., Lei, H., Meng, X., Wang, M. (eds.) ICEBE, pp. 535–540. IEEE Computer Society, Los Alamitos (2005)
2. Markovic, I., Pereira, A.C., de Francisco, D., Munoz, H.: Querying in business process modeling. In: SeMSoC: Proceedings of the 2nd International Workshop on Business Oriented Aspects concerning Semantics and Methodologies in Service-oriented Computing, Vienna, Austria, 2007 (to appear)
3. Sangiorgi: A theory of bisimulation for the pi-calculus. ACTAINF: Acta Informatica 33 (1996)
4. Curtis, B., Kellner, M.I., Over, J.: Process modeling. Comm. of the ACM 35(9), 75–90 (1992)
5. Jablonski, S., Bussler, C.: Workflow Management: Modeling Concepts, Architecture, and Implementation. International Thomson Computer Press, London, UK (1996)
6. Markovic, I., Pereira, A.C.: Towards a formal framework for reuse in business process modeling. In: semantics4ws: Proceedings of the 2nd International Workshop on Advances in Semantics for Web services, Brisbane, Australia, 2007 (to appear)

[3] www.ip-super.org

7. Puhlmann, F.: Why do we actually need the pi-calculus for business process management? In: Abramowicz, W., Mayr, H.C. (eds.) BIS, LNI, GI, vol. 85, pp. 77–89 (2006)
8. Puhlmann, F., Weske, M.: Using the pi-calculus for formalizing workflow patterns. In: Business Process Management, pp. 153–168 (2005)
9. Overdick, H., Puhlmann, F., Weske, M.: Towards a formal model for agile service discovery and integration. In: International Workshop on Dynamic Web Processes, associated with the 3rd International Conference on Service Oriented Computing, Amsterdam, The Netherlands (2005)
10. Puhlmann, F.: A tool chain for lazy soundness. In: Business Process Management, pp. 9–16 (2006)
11. Anja Bog, F.P.: A tool for the simulation of pi-calculus systems. In: Open.BPM 2006: Geschäftsprozessmanagement mit Open Source-Technologien, Hamburg, Germany (2006)
12. Yu, E.S.K., Mylopoulos, J.: Understanding "why" in software process modeling, analysis, and design. In: Proceedings of the 16th International Conference on Software Engineering, pp. 159–168. IEEE Computer Society Press, Los Alamitos (1994)
13. Kifer, Lausen, Wu: Logical foundations of object-oriented and frame-based languages. JACM: Journal of the ACM 42 (1995)
14. BPMI: Business Process Query Language, http://www.service-architecture.com/web-services/articles/business_process_query_language_bpql.html
15. Beeri, C., Eyal, A., Kamenkovich, S., Milo, T.: Querying business processes. In: Dayal, U., Whang, K.Y., Lomet, D.B., Alonso, G., Lohman, G.M., Kersten, M.L., Cha, S.K., Kim, Y.K. (eds.) VLDB, pp. 343–354. ACM, New York (2006)
16. Mou, Y., Cao, J., sheng Zhang, S.: A process component model for enterprise business knowledge reuse. In: IEEE SCC, pp. 409–412. IEEE Computer Society, Los Alamitos (2004)
17. Victor, B., Moller, F.: The Mobility Workbench — a tool for the π-calculus. In: Dill, D.L. (ed.) CAV 1994. LNCS, vol. 818, pp. 428–440. Springer, Heidelberg (1994)

Prototypical Implementation of a Pragmatic Approach to Semantic Web Service Discovery during Process Execution

Christian Stamber*, Sebastian Stein*, and Marwane El Kharbili

IDS Scheer AG, ARIS Research, Altenkesseler Str. 17, D-66115 Saarbrücken, Germany
{christian.stamber,sebastian.stein,marwane.elkharbili}@ids-scheer.com
http://www.ids-scheer.com/

Abstract. The usage of semantic web services in enterprise computing promises many advantages. Instead of hard-wiring web services in business processes during design-time, business experts just specify the expected functionality and matching semantic web services are bound automatically during runtime. However, the introduction of semantic web services requires the replacement or extension of existing middleware and standards. We suggest a more realistic approach based on existing standards like BPEL. We provide a prototypical implementation as a proof of concept. Our solution helps to accelerate the adoption of semantic technologies in business process management and provides researchers with early feedback.

Keywords: business process automation, semantic web service, BPEL, web service discovery.

1 Introduction

Business processes are a core asset of every company. They define the dynamic interaction of the elements constituting the company like employees, machines, partners, and resources. As the actual products become more and more commodities, the business processes delivering the products must provide the competitive advantage so that a company can survive. This includes adapting the business processes quickly to a changing business environment.

As a matter of fact, business processes are supported today by information technology (IT). Even though IT might not provide any competitive advantage on its own [1], having no IT support at all will certainly lead to failure. Therefore, it is the aim of business process automation to derive an IT implementation out of the business processes. A straight-forward solution is prevented, because of the existence of the business-IT divide [2]. Business experts usually have a social science background whereas IT experts have a natural science background and

* Both authors contributed equally to this paper.

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 201–212, 2008.

therefore both groups use different terms and tools. For example, business experts use process modelling languages like the Event-driven Process Chain (EPC) notation [3] or the Business Process Modeling Notation [4] to define and document the business processes. On the other hand, IT experts require executable process descriptions in a process language like BPEL [5]. The business process model of the business expert is not as formal as the executable model of the IT expert. Information relevant for execution like technical exception handling is missing and cannot be added by business experts because of lacking knowledge. In addition, by taking a look at the paradigm of model driven architecture [6] it can be seen that business processes are usually defined on a platform independent level whereas executable process models are defined platform specific. Adding platform specific details to a platform independent business process is no option.

According to Leymann [7] business processes implemented by orchestrated web services is the future direction for business application development. As of today, web services to be used for automating a business function are specified during design-time of the business process model. Stein and Ivanov [8] discuss a transformation how an executable BPEL model can be derived from an EPC model. The major drawback of this approach lies in the fact that appropriate web services must be selected during design-time. In a related paper, Stein et al. [9] show how selecting a web service during design-time by business experts can be supported. Still, it does not allow a flexible reconfiguration of the web service orchestration during runtime, because the web services are hard-coded in the executable process model.

1.1 Problem Statement

Hepp et al. [10] envision the usage of semantic technologies in enterprise computing to overcome current problems, especially bridging the business-IT divide. In their approach, a business expert does not select a specific web service anymore during business process modelling but instead just specifies the required functionality. This semantic requirements specification is used during execution of the process model to discover and bind web services. Their approach is based on the assumption that web services are available as semantic web services [11], so that the semantic descriptions given in the process model can be matched against them. Hepp et al. [12] propose a stack of ontologies to be used to capture the business and IT knowledge at different abstraction levels. The ontology stack uses the Web Service Modeling Ontology (WSMO) [13] to define semantic web services.

If following the approach taken by Hepp et al. [10,12], a heavily extended infrastructure is needed. For example, the business process modelling tool used by the business experts must be able to serialise the process model in the Business Process Modeling Ontology (BPMO) [12, see]. Even though preliminary work on such a modelling tool exists [14], based on a semantically extended BPMN notation [15], it will take several years of implementation. The same is true for a BPEL process execution engine able to execute a semantically extended BPEL version. Nitzsche et al. [16] defined an ontology based on BPEL, which can be used as a starting point for such a semantic execution engine.

This massive upfront investment in a semantically extended IT landscape hinders the adoption of semantic technologies into action. This prevents companies to leverage the advantages of semantic technologies and on the other hand researchers do not get enough early feedback on their work. To overcome this problem, we contribute a pragmatic solution allowing the discovery of web services during runtime based on semantic descriptions. There are a few constraints we considered while designing and implementing the solution:

- It must be possible to reuse existing business process models, because companies already invested heavily in business process modelling. Those investments must be preserved by reusing existing artefacts.
- Creating the executable BPEL model must be automated.
- The proposed solution must be seamlessly integrated into the existing tool chain.
- The executable process model must be defined in standard BPEL version 1.1, so that it can be executed on any existing process engine supporting this standard.
- The solution must be independent of the semantic formalism used. For example, it should be possible to describe the web services semantically using WSMO [13] or OWL-S [17].
- The solution must be extensible so that different semantic environments like WSMX [18] and IRS III[19] can be supported, if needed.

The following sub-section discusses our research design to answer the research question under the given constraints.

1.2 Research Design

Implementing a solution for semantic service discovery during process execution by using existing standards and without changing the tool chain is challenging. We[1] therefore defined an iterative research design to refine our solution. The main phases are as follows:

1. Define an initial solution by investigating existing standards, tools, and techniques.
2. Implement a prototype to validate the general feasibility of the solution.
3. Use a real-world use-case to validate the feasibility of the solution in a realistic setting.
4. Conduct an empirical study [20, see e. g.] to show the advantage of semantic business process automation compared to current technologies.

The research effort is still work in progress. In this paper we report on the results of the first two steps in our research design. We have defined our solution

[1] The presented research is part of the SUPER research project (http://www.ip-super.org/), supported by the European Union's 6th Framework Programme.

and validated it by implementing a prototype using a trivial business process. We are currently collaborating with an industrial partner from the telecommunication domain to define a more realistic use-case. We will use this use-case to demonstrate the feasibility of our approach in a real-world scenario.

2 Solution

In this section we present the solution. In our prototype, we use IDS Scheer's ARIS SOA Architect[2] as modelling tool. Oracle's SOA Suite[3] and especially the Oracle BPEL Process Server are used to execute the generated BPEL process. We use WSMX [18] for semantic web service discovery during runtime. We developed a small integration component to connect the Oracle BPEL Process Server and WSMX. This integration component is called Semantic Invocation Service (SISi) and is described in detail in sub-section 2.3. It is a Java servlet using Apache Axis2[4] for web service support. SISi is deployed on the Java servlet container Apache Tomcat[5]. We use WSMO [13] for semantic descriptions.

The next sub-section presents an overview of our general approach.

2.1 Overview

The general approach is illustrated as a simplified EPC process model in figure 1. The process consists of seven steps executed by different experts. Step four, six and seven have a special symbol in the process model to illustrate that they can be mostly automated.

In step one, we use a domain ontology created before to describe a set of existing web services semantically. We do not inject the semantic description into the existing syntactical web service interface descriptions given as WSDL [21] files, even though there is an official recommendation called SA-WSDL [22] to do that. SA-WSDL is not supported yet by many tools, so we decided to use a less intrusive way. We create a WSMO description for each web service using WSMO Studio[6] [23]. The WSMO description defines the capabilities of the web service and provides a pointer to the belonging WSDL web service. This is a satisfying solution for our prototypical implementation.

In step two, the business process is modelled. We use ARIS SOA Architect for this purpose, which supports the EPC as well as the BPMN notation. We have decided to use the EPC notation, because ARIS SOA Architect provides a transformation of EPC to BPEL, which can be customized.

In step three, we annotate the business process model semantically. As we use WSMO, we specify a WSMO Goal for each business function in the EPC. A WSMO Goal [13, see for details] captures what capabilities are required. We use

[2] http://www.aris.com/soa/
[3] http://www.oracle.com/technologies/soa/soa-suite.html
[4] http://ws.apache.org/axis2/
[5] http://tomcat.apache.org/
[6] http://www.wsmostudio.org/

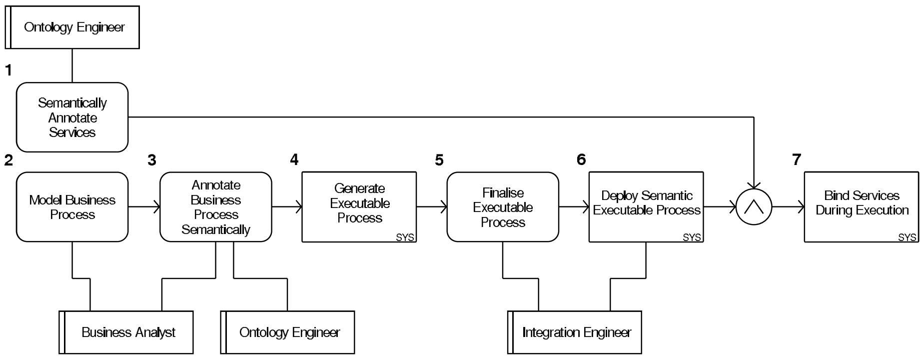


Fig. 1. Overall Approach

WSMO Studio for the definition of the WSMO Goals. We have implemented a script allowing us to import WSMO Goals into ARIS. A WSMO Goal is mapped in ARIS to the capability object type. This is a perfect fit, because the capability object type can be related to business functions to express required functionality. The capability object type is given the name of the WSMO Goal and it has a user defined attribute pointing to the actual WSMO definition.

In step four, we automatically generate an executable BPEL process using ARIS SOA Architect's built-in EPC to BPEL transformation. To also support the transformation of the semantic annotations, we extended the transformation slightly. The result of this step is a graphical representation of the generated BPEL process.

In step five, we finalise the BPEL process and export it to the XML representation. We use the semantic invocation service (SISi) to inject the semantic information in a standard conform way. This is explained in detail in sub-section 2.3.

In step six, we deploy the BPEL process on the Oracle BPEL Process Server. We prepare WSMX to answer semantic discovery requests and we deploy SISi on Apache Tomcat. For the prototype, we have automated this step with some custom scripts.

In step seven, the BPEL process is executed. Each time a semantic discovery must be done, the semantic request is forwarded to SISi. SISi first calls WSMX to discover a matching web service. Afterwards, SISi invokes the discovered web service and returns its output to the calling BPEL process.

2.2 Execution of a Semantic Process

We use the BPEL language version 1.1 as format for executable business processes. BPEL allows orchestrating a set of web services and there are many middleware products by various vendors supporting this standard. BPEL itself has a mechanism to support dynamic binding during runtime. Each web service is represented as a partnerlink in BPEL. The partnerlinks are a special kind of variables specifying the web service to be called. It is possible to exchange the content of a partnerlink during runtime by assigning a new value to it. However, it can only be exchanged with content of the same partnerlink type. Because of this limitation, we decided to not use this mechanism of BPEL. Instead, we introduce the usage of a semantic invocation service (SISi). Each time a web service should be discovered during runtime, SISi is called. The following sub-section explains in detail the architecture and interface design of SISi.

2.3 Semantic Invocation Service (SISi)

A central component of our solution to semantic web service discovery during process execution is the semantic invocation service (SISi). Figure 2 sketches the architecture. It can be seen that we defined a classical layered software

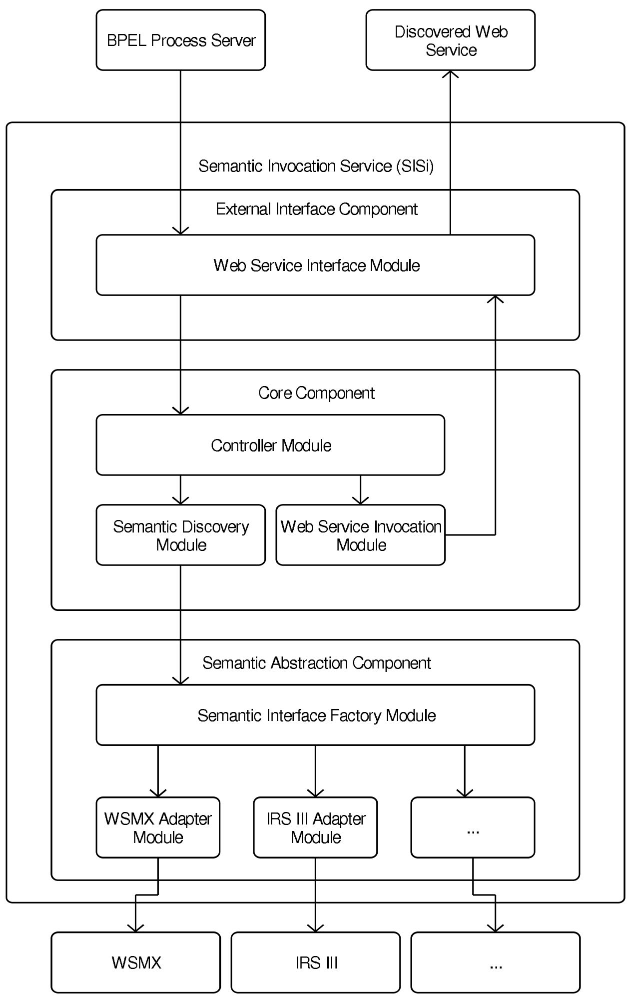


Fig. 2. Architecture of Semantic Invocation Service (SISi)

architecture [24, see e. g.] consisting of three layers. We provide an OpenSource reference implementation of SISi[7] licensed under the GNU Lesser General Public License version 2.1[8]. This licence allows the usage of SISi in commercial as well as in OpenSource projects.

The top layer consists of the *External Interface Component* exposing SISi's functionality to external users. Currently, we only provide a *Web Service Interface Module*. This module is mostly generated code based on the Apache Axis2 web service framework. The web service interface is used by the BPEL process server to invoke SISi. SISi itself uses the web service interface module to invoke the discovered web service.

The actual application logic of SISi is available in the *Core Component*. A central *Controller Module* receives the semantic service discovery request and uses the *Semantic Discovery Module* to initiate the discovery. In a second step, the controller module uses the *Web Service Invocation Module* to call the discovered web service. Both modules used by the controller are rather small in the current implementation. They mainly forward the requests to the other components of SISi. We decided to include the modules anyway to ensure extensibility to fulfil future requirements. For example, if data mediation is needed in a future version, this can be added to the semantic discovery module without having to change the controller.

The bottom layer called *Semantic Abstraction Component* provides access to the semantic discovery components. Even though there are preliminary efforts[9] to standardise such components in a reference architecture, we are not aware of any widely accepted standard. Therefore, an individual adapter module is required for each semantic discovery component to be supported. The current implementation contains only an adapter module for WSMX. All adapter modules must implement the same set of interface operations so that they can be used transparently through the *Semantic Interface Factory Module*. This approach allows us great flexibility, because any specifics of the different semantic discovery components to be supported are pushed into a single module.

Figure 3 shows the input consumed and the output produced by SISi. In order to perform its task, SISi needs the semantic description, which is a WSMO Goal in our case, and it needs the input message for the web service to be invoked. As a result, SISi returns the message it received from the invoked web service. As SISi can't foresee, which web service will be found and invoked, it can't provide an operation with parameters as the discovered web service has.

Code snippet 4 shows the WSDL definition of SISi and code snippet 5 shows the belonging data definition of the different message parts. It can be seen that the input message consists of two parts – the semantic description and the message to be forwarded to the discovered web service. Currently, the type of the message part for the semantic description is just a plain string. On the other hand, we do not know what type is needed for the input message of the

[7] http://code.google.com/p/semanticinvocationservice/
[8] http://www.gnu.org/licenses/old-licenses/lgpl-2.1.html
[9] http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=semantic-ex

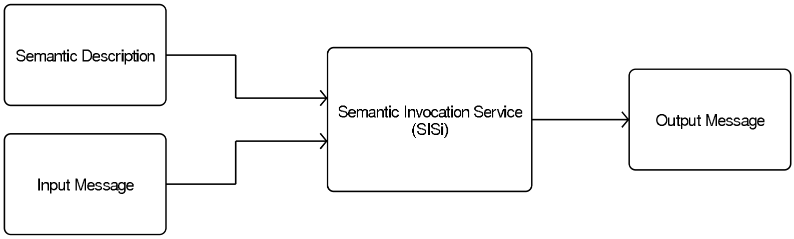


Fig. 3. Input Consumed and Output Produced by SISi

```
<?xml version="1.0" encoding="utf-8"?>
<wsdl:definitions name="WebServiceInterfaceWS"
 targetNamespace="http://sisi.externalInterface/"
 xmlns:wsdl="http://schemas.xmlsoap.org/wsdl/"
 xmlns:tns="http://sisi.externalInterface/"
 xmlns:dataNs="http://sisi.externalInterface/dataTypes"
 xmlns:xsd="http://www.w3.org/2001/XMLSchema">
 <wsdl:types>
  <schema xmlns="http://www.w3.org/2001/XMLSchema">
   <import namespace="http://sisi.externalInterface/dataTypes"
     schemaLocation="SISi_WebServiceInterface_dataTypes.xsd"/>
  </schema>
 </wsdl:types>
 <wsdl:message name="invokeSemanticWebServiceRequest">
  <wsdl:part name="semanticDescription" element="xsd:string"/>
  <wsdl:part name="parameters" element="dataNs:hashMap"/>
 </wsdl:message>
 <wsdl:message name="invokeSemanticWebServiceResponse">
  <wsdl:part name="parameters" element="dataNs:hashMap"/>
 </wsdl:message>
 <wsdl:portType name="WebServiceInterfaceWS">
  <wsdl:operation name="invokeSemanticWebService">
   <wsdl:input message="invokeSemanticWebServiceRequest"/>
   <wsdl:output message="invokeSemanticWebServiceResponse"/>
  </wsdl:operation>
 </wsdl:portType>
</wsdl:definitions>
```

Fig. 4. WSDL Definition of the Semantic Invocation Service (SISi)

discovered web service. Therefore, we use a hash map, which can contain objects of any type. The output message of SISi just consists of one message part. This message part transports the message received from the discovered and invoked web service back to the calling BPEL process. Again, as we do not know the format of this message, we also use a hash map able to store a collection of objects of any type.

```
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema"
 xmlns:tns="http://sisi.externalInterface/dataTypes">
 <xsd:complexType name="hashMap">
  <xsd:complexContent>
   <xsd:extension base="map">
    <xsd:sequence/>
   </xsd:extension>
  </xsd:complexContent>
 </xsd:complexType>
 <xsd:complexType name="map">
  <xsd:sequence>
   <xsd:element name="mapEntry" type="mapEntry" minOccurs="0"
    maxOccurs="unbounded"/>
  </xsd:sequence>
 </xsd:complexType>
 <xsd:complexType name="mapEntry">
  <xsd:sequence>
   <xsd:element name="key" type="xsd:anyType"/>
   <xsd:element name="value" type="xsd:anyType"/>
  </xsd:sequence>
 </xsd:complexType>
</xsd:schema>
```

Fig. 5. Data Definition of the Semantic Invocation Service (SISi)

3 Discussion

Our solution enables us to do semantic web service discovery during process execution using standard middleware. We achieve that by providing a semantic invocation service (SISi), which is a standard conform WSDL web service. SISi forwards the semantic request to a semantic discovery environment and invokes the discovered web service. So far, the prototypical implementation of SISi only supports WSMX as semantic environment. Even though SISi was designed in a way to be easily extended with support for other semantic environments, it might turn out that changes to SISi are needed. For example, at the current point SISi only exposes one operation, which does not allow selecting the semantic execution environment to use. However, we are convinced that the current implementation is a valid proof of concept showing that our approach works.

In general, we must make sure to not make SISi too complex. Each proposed extension for SISi must be carefully reviewed and it must be checked if it is not possible to push the required functionality down to the semantic environments. It is the purpose of SISi to act as a proxy between the BPEL process and the semantic environment, but SISi is not meant to be a semantic environment on its own. At some point it will be better to directly integrate semantic functionality into existing middleware. SISi is just meant as a solution for the time till that happens.

The current prototypical implementation and our approach have some shortcomings. For example, currently we are only able to handle synchronous web service requests. This is a major disadvantage, because BPEL is meant as a language for long-running processes, which usually consist of asynchronous web service calls. SISi also doesn't throw an exception in case it can't find a semantic web service. However, adding such extensions is relatively straight-forward. Another problem area is SISi's web service interface. The input message type uses a hash map able to store any kind of object. This is necessary, because the type of the received message cannot be predicted. However, having such a broad message definition prevents to use type checking techniques. It is also impossible to apply any validation techniques on the input received. This might cause security problems, if malformed requests are received. Also, denial of service attacks are possible, because the caller can send an endless stream of objects in the hash map. However, we are ignoring the security problems at the moment, because SISi is not meant for production usage, but instead as a prototype to foster research. Still, SISi will allow researchers to conduct more realistic case studies, because they don't have to reinvent the whole middleware stack.

In order to inject semantic web service discovery in the BPEL process, we have to replace each discovery request by an invocation of SISi. This makes the BPEL process unreadable for human users. We do not consider that as a major drawback, because the real editing should happen on the business process level and not on the BPEL process.

SISi supports only semantic service discovery at the moment, but a natural next step would be to also use data mediation. This would require expressing the input and output data as ontologies as well, which increases the necessary modelling effort. This also has implications how business process modelling is done and might require a further extension of business process modelling tools. On the other hand, it might simplify the generation of BPEL, because a transformation does not have to take care of data handling anymore, because data handling is done by an intelligent mediation component. We have decided to not include semantic data mediation in the first version to make the prototype not too complex.

Our solution is independent of the semantic formalism used and it can also be integrated with different semantic discovery environments. This is a major advantage, because it also allows comparing different semantic formalisms and discovery environments. SISi is provided as an OpenSource implementation to the public. If interested, other researchers and practitioners are able to extend SISi so that it supports other semantic formalisms and discovery environments.

The main advantage of our approach is that it can be used already today. There is no need to replace major parts of the enterprise computing stack or to heavily invest in new middleware. This frees up much needed resources, which can be used to do more fascinating research in context of semantic business process management in contrast of implementing middleware software again. Practitioners as well as researchers can test the usage of semantic technologies already today, gather feedback and provide a test environment for new semantic research insights. We are convinced that this will accelerate the adoption of semantic technologies in enterprise computing and also stimulate research focused on enterprise needs.

References

1. Carr, N.G.: Does IT matter? Information Technology and the Corrosion of Competitive Advantage. Harvard Business School Press (2004)
2. Smith, H., Fingar, P.: Business Process Management: The Third Wave, 1st edn. Meghan-Kiffer Press, Tampa, FL, USA (2003)
3. Scheer, A.W., Thomas, O., Adam, O.: Process modelling using event-driven process chains. In: Dumas, M., van der Aalst, W.M.P., ter Hofstede, A.H.M. (eds.) Process-Aware Information Systems, pp. 119–146. Wiley, Hoboken, New Jersey, USA (2005)
4. OMG: Business process modeling notation (bpmn) specification. Technical report, Object Management Group (OMG) (February 2006), http://www.omg.org/docs/dtc/06-02-01.pdf
5. Andrews, T., Curbera, F., Dholakia, H., Goland, Y., Klein, J., Leymann, F., Liu, K., Roller, D., Smith, D., Thatte, S., Trickovic, I., Weerawarana, S.: Business process execution language for web services (bpel4ws) 1.1. Technical report (May 2003), http://www-128.ibm.com/developerworks/library/ws-bpel/
6. Miller, J., Mukerji, J.: MDA guide. Technical Report omg/2003-06-01, Object Management Group (OMG), Version 1.0.1 (June 2003)
7. Leymann, F.: The influence of web services on software: Potentials and tasks. In: 34th Annual Meeting of the German Computer Society, Ulm, Germany, September 20-24, 2004, Springer, Heidelberg (2004)
8. Stein, S., Ivanov, K.: EPK nach BPEL Transformation als Voraussetzung für praktische Umsetzung einer SOA. In: Bleek, W.G., Raasch, J., Züllighoven, H. (eds.) Software Engineering 2007, Hamburg, Germany, Gesellschaft für Informatik (GI), March 2007. Lecture Notes in Informatics (LNI), vol. 105, pp. 75–80 (2007)
9. Stein, S., Barchewitz, K., El Kharbili, M.: Enabling business experts to discover web services for business process automation. In: Pautasso, C., Gschwind, T. (eds.) 2nd Workshop on Emerging Web Services Technology, Halle, Germany, November 2007, pp. 19–35 (2007)
10. Hepp, M., Leymann, F., Domingue, J., Wahler, A., Fensel, D.: Semantic business process management: A vision towards using semantic web services for business process management. In: IEEE International Conference on e-Business Engineering (ICEBE 2005), Beijing, China, pp. 535–540 (2005)
11. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American 284(5), 34–44 (2001)
12. Hepp, M., Roman, D.: An ontology framework for semantic business process management. In: 8th International Conference Wirtschaftsinformatik 2007, Karlsruhe, pp. 423–440 (2007)
13. Fensel, D., Lausen, H., Polleres, A., de Bruijn, J., Stollberg, M., Roman, D., Domingue, J.: Enabling Semantic Web Services: The Web Service Modeling Ontology. Springer, Heidelberg (2006)
14. Dimitrov, M., Simov, A., Stein, S., Konstantinov, M.: A BPMO based semantic business process modelling environment. In: Hepp, M., Hinkelmann, K., Karagiannis, D., Klein, R., Stojanovic, N. (eds.) Workshop on Semantic Business Process and Product Lifecycle Management (SBPM 2007), CEUR Workshop Proceedings, Innsbruck, Austria, June 2007, Vol. 251, pp. 101–104 (2007)
15. Abramowicz, W., Filipowska, A., Kaczmarek, M., Kaczmarek, T.: Semantically enhanced business process modelling notation. In: Hepp, M., Hinkelmann, K., Karagiannis, D., Klein, R., Stojanovic, N. (eds.) Workshop on Semantic Business Process and Product Lifecycle Management (SBPM 2007), CEUR Workshop Proceedings, Innsbruck, Austria, June 2007, vol. 251, pp. 88–91 (2007)

16. Nitzsche, J., Wutke, D., van Lessen, T.: An ontology for executable business processes. In: Hepp, M., Hinkelmann, K., Karagiannis, D., Klein, R., Stojanovic, N. (eds.) Workshop on Semantic Business Process and Product Lifecycle Management (SBPM 2007), CEUR Workshop Proceedings, Innsbruck, Austria, June 2007, vol. 251, pp. 52–63 (2007)
17. Martin, D., Burstein, M., Hobbs, J., Lassila, O., McDermott, D., McIlraith, S., Narayanan, S., Paolucci, M., Parsia, B., Payne, T., Sirin, E., Srinivasan, N., Sycara, K.: Owl-s: Semantic markup for web services. Technical report (2004), http://www.daml.org/services/owl-s/
18. Haller, A., Cimpian, E., Mocan, A., Oren, E., Bussler, C.: Wsmx - a semantic service-oriented architecture. In: International Conference on Web Service (ICWS 2005), Orlando, Florida, USA (2005)
19. Domingue, J., Cabral, L., Hakimpour, F., Sell, D., Motta, E.: Irs iii: A platform and infrastructure for creating wsmo based semantic web services. In: Workshop on WSMO Implementations (WIW 2004), Frankfurt, Germany (2004)
20. Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A.: Experimentation in software engineering: an introduction. In: International Series in Software Engineering, Kluwer Academic Publishers, Norwell, USA (2000)
21. Christensen, E., Curbera, F., Meredith, G., Weerawarana, S.: Web service description language (wsdl) 1.1. Technical report, W3 Consortium (March 2001), http://www.w3.org/TR/wsdl
22. Farrell, J., Lausen, H.: Semantic annotations for wsdl. Technical report, W3 Consortium (2007), http://www.w3.org/TR/sawsdl/
23. Dimitrov, M., Simov, A., Konstantinov, M., Momtchev, V.: Wsmo studio - a semantic web services modelling environment for wsmo (system description). In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 749–758. Springer, Heidelberg (2007)
24. Hofmeister, C., Nord, R., Soni, D.: Applied software architecture, 2nd edn. Addison-Wesley, Massachusetts (2000)

SAF: Strategic Alignment Framework for Monitoring Organizations

Ernesto Damiani[1], Fabio Mulazzani[2], Barbara Russo[2], and Giancarlo Succi[2]

[1] Università di Milano - Dipartimento di Tecnologie dell'Informazione
Via Bramante, 65 - Crema - Italy
damiani@dti.unimi.it

[2] Libera Università di Bolzano - Faculty of Computer Science
Piazza dei Domenicani, 3 - Bolzano/Bozen - Italy
{fabio.mulazzani,barbara.russo,giancarlo.succi}@unibz.it

Abstract. Reaching a Strategic Alignment is a crucial aspect for any organization. The alignment can be achieved by controlling, through monitoring probes, the coherency of the Business Processes with the related Business Strategy. In this paper we present SAF, a powerful framework for those organizations that aim at a superior business performance and want to keep monitored the organization's alignment. SAF has been applied to a real case study and it has also been compared with GQM$^+$ Strategy [2] and Process Performance Indicators Monitoring Model [16].

Keywords: Business Process Modeling, Business Strategy, Strategic Alignment, Process Monitoring, Workflow Probe, Alignment Framework.

1 Introduction

Any organization needs to be supported by an infrastructure IT in order to accomplish its business mission [22]. The development and the adoption of an IT Governance that orchestrates all the IT potentials can lead to superior performance of the organization. A crucial requirements for IT Governance is aligning or defining a set of operational goals and an operational management that allow the mapping between the Business Strategy (BS) and the Business Processes (BPs), this is strategic alignment that is defined by Reich and Benbasat in [19] as the degree to which the IT mission, objectives, and plans support and are supported by the business mission, objectives, and plans. In fact, according to [2] an organizational management system will not guarantee the organizational success unless the BS is translated into a set of operational goals and a quantitative business management.

The manager issues of IT Governance related to strategic alignment are:

– **Performance measurement** - This is an activity that tracks and monitors: strategy implementation, project completion, resource usage, process performance and service delivery.

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 213–226, 2008.

– **Control and Accountability** - This is about the existence and practical application of a proper control and accountability system.

In this paper we present a Strategic Alignment Framework (SAF) that combines the expressivity power of Jackson Problem Frames [12] together with standardized Object Management Group (OMG) ontology and modeling languages. For measuring the level of goal achievement in business processes, solutions able to constantly monitor the workflow content and performance have to be deployed over it. Unfortunately, a methodology for realizing and representing the required monitoring - including the instrumentation - does not exist yet [16]. SAF also provides controls over the workflow for the BP performance, these are called probes. The framework presented has been compared with other two alignment models as described in Section 4. SAF has also been applied over a well known ficticius case of a car rental company as initial validation (already described in [7]), and over a real case of an italian electric company.

The paper is organized as follows. Section 2 provides an overview of the related works on both alignment framework and on process monitoring. Section 3 describes the framework. Section 4 describes the comparison with the related models. Section 5 presents the real case study applied to the framework. Section 6 concludes the paper.

2 Related Works

There exists various techniques to verify the alignment between the three main levels of an organization: the strategy, the business processes, and the Information System (IS). In this section, we briefly summarize the literature on the strategic alignment, on the goal modeling, and on the process monitoring.

In [4] Bleistein et al. adopt i^* - a goal modeling notation for information systems requirements - [23] to represent the strategy and the context diagram from Jacksons Problem Frames to model the IT context of strategy achievement. As stated in [21] the approach in [4] represents the business strategies and the IS description in the same model, and with the same formalism. Furthermore it relates BS and IS description through simple and mono-typed contribution links. One issue with this approach relates to the i^* notation, which does not use the black box/white box strategy and produces complex models when the number of goals increases. Indeed the main difficulties with i^* are that (i) it lacks of a systematic goal refinement mechanisms, and (ii) it has no goal-strategy coupling to help clarifying the multiple ways in which a goal can be achieved.

In [11] is presented $e^3{-}Value$, a modeling notation for e-commerce systems. It enables validation of requirements in terms of the system's potential to generate economic value. However, $e^3{-}Value$ has been recognized as ignoring key elements of business value analysis derived by the business strategy [17,8,18].

In [10] the authors defined two main groups of metrics corresponding to different facets of the process-to-organization alignment, one for generic metrics, and the other for model-specific ones.

Much research has also been done on the definition and properties of business process performance indicators. In the early Nineties, Kaplan and Norton introduced the notion of a *Balanced Scorecard* [14] as a set of indicators capable of dealing with business goals related to all the four *dimensions* relevant to monitoring business organizations (Financial Requirements, Customer Relations, Internal Business Processes, Learning and Growth Capabilities). Most approaches agree that process-level indicators must be chosen by balancing the coverage of business goals with the data collection and processing costs.

Other works from the software engineering research community, focus on the alignment between business modeling and software development or execution. As discussed in [13] the general idea is that requirements models are built from organizational goals. Some of these works, such as for instance [15] or [9], are strongly related to UML and do not clearly specify some organizational aspects such as technology that implements business processes or the relationships among the different organizational views.

In the following two subsections we will describe the two models considered for the comparison in Section 4.

2.1 GQM$^+$ Strategy Method

The Goal Question Metric (GQM) approach [1] provides a top-down paradigm for an organization or a project to define goals, refine those goals down to specifications of data to be collected, and then analyze and interpret the resulting data with respect to the original goals. GQM goals are defined in terms of purpose, focus, object of study, viewpoint, and context. Such a goals are then refined into specific questions that must be answered in order to evaluate the achievement of the goal. The questions are then operationalized into specific quantitative measures. The GQM formalizes the deduction process that derives the appropriate measures to answer a given business goal.

However GQM has never provided explicit support for integrating its software measurement model with elements of the larger organization, such as higher-level business goals, strategies, and assumptions [3]. That is why Basili et al. in [2] propose and describe a method that adds several extensions on top of the GQM model. The GQM$^+$ Strategy method makes the business goals, strategies, and corresponding software goals explicit. The GQM$^+$ Strategy method also makes the relationships between software-related activities and measurement goals explicit. Sequence of activity necessary for accomplishing the goals are defined by the software organization and embedded into scenarios in order to achieve some software-related goal. Links are established between each software goal and the business-level strategy it supports. Attached to goals, strategies, and scenario at each level of the model is information about relationships between goals, relevant context factors, and assumptions.

2.2 Process Performance Indicators Monitoring Model

In [16] Cristof Momm et al. present a model-driven methodology for a top-down development of a process-oriented IT support based on Service Oriented

Architecture (SOA). In contrary to existing approaches Momm also include the monitoring required for business process controlling and introduce metamodels for the specification of process performance indicators (PPIs) in conjunction with the necessary monitoring. As mentioned in [16], a PPI is attached to the concept Process as part of the computation-independent process model. Optionally a TargetValue indicating the objective as well as an AlarmValue defining a threshold for an intervention may be specified. The PPI is further characterized by assigning a Dimension. Thereby information like the data type, the direction (e.g. ascending or descending) and the unit of the value are specified. The calculation of the mandatory CurrentValue on the basis of runtime information provided by the underlying orchestration is handled by the PPIMonitor. This aspect is tackled within the scope of the PPI monitoring model. Furthermore, they have distinguished between basic and aggregated PPIs.

3 The Framework

Business processes are deployed and executed in evolving organizational and technological environments. SAF (see Fig. 1) is a powerful tool for those organizations that need to:

- align - make coherent - their BPs together with their BSs;
- align their IT infrastructure together with the existing BPs;
- create an IT infrastructure to support a newly founded business organization.

We focus on the first point.

In order to represent the environment where the business organization activities take place, we use an approach inspired by Problem Frames (PF)[12] (see Subsection 3.2), originally introduced by Michael Jackson as a technique for representing software requirements. We then combined the previous technique together with the Business Motivation Model(BMM) (see Section 3.1) to elicitate the strategy that generates the environment's requirements in the PF. The link between the BMM and the Business Process Diagram (BPD), that in our framework is included inside a PF, is justified by the element *Course of Action* that represents the overall approach that the enterprise will take to achieve its *Desired Results*. *Course of Action* can be realized by, or made operative by, BP as explained in [5].

The most important innovation of our framework is the adoption, hence the modeling, of probes (see Section 4.2) that operates on the IT to monitor the activities taken in action by a task, or a group of tasks, of a specific process. This is to control the achievement of the strategies that have generated the controlled process. Probes consist of a memory for BP allocation and a logic for the BP monitoring. So far, the logic modeling and the deployment of the probes over the BPD has been done using the expertize as mentioned in Section 4.2 and 6. This approach may result hard to automatize. Our future goal will be to define models and rules (logic and deployment) of probes to facilitate their implementation in a software platform supporting SAF.

The external and internal measurement of the BPs, is essential to verify the achievement of business strategic goals and, in case they are not completely achieved, to enable the management to change the strategy that generated the monitored process.

The use of PF - the lower component of Fig. 1 - has been contextualized in this framework by modeling the elements of the *Domain Context* with the Semantics of Business Vocabulary and Business Rules (SBVR - See Subsection 3.2), and the requirements with a BPD in Business Process Modeling Notation (BPMN). The advantage of using SBVR is that it can exhaustively describes the entities of the context and their relations. The use of BPD in the requirements, generated by the *Course of Action* of the BMM, represents the procedures used by the entities of the Domain Context.

In Subsection 3.1 we summarize the basic concept of the BMM while in Subsection 3.2 we present the Jackson PF approach and its business modeling contextualization operated by SBVR and BPMN. The monitoring probes are descived in Section 4.2 after the comparison of SAF with other two models.

3.1 Business Motivation Model (BMM)

The Business Motivation Model (BMM)[5], according to its developers, the Business Rules Group, is a metamodel of the concepts essential for business governance. The BMM provides:

- A vocabulary for governance including such concepts as "influencer", "assessment", "business policy", "strategy", "tactic", "goal", and fact type that relate them, such as "business policy governs course of action".
- Implicit support for an end-to-end process that runs:
  – From recognition that an influencer (regulation, competition, environment, etc) has an impact on the business;
  – To implementing the reaction to that impact in business processes, business rules and organization responsibilities.
- The basis for logical design of a repository for storage of BMMs for individual businesses.

There are two major components of the BMM.

- The first is the Ends and Means of business plans. Among the Ends are things the enterprise wishes to achieve - for example, Goals and Objectives. Among the Means are things the enterprise will employ to achieve those Ends - for example, Strategies, Tactics, Business Policies, and Business Rules.
- The second is the Influencer that shape the elements of the business plans, and the Assessments made about the impacts of such Influencers on Ends and Means (i.e. Strengths, Weaknesses, Opportunities, and Threats).

All elements of the BMM are developed from a business perspective. The basic idea is to develop a business model for the elements of the business plans before system design or technical development is begun. In this manner, the business plans can become the foundation for such activity, connecting system solutions firmly to their business intent. i^* notation does not provide this.

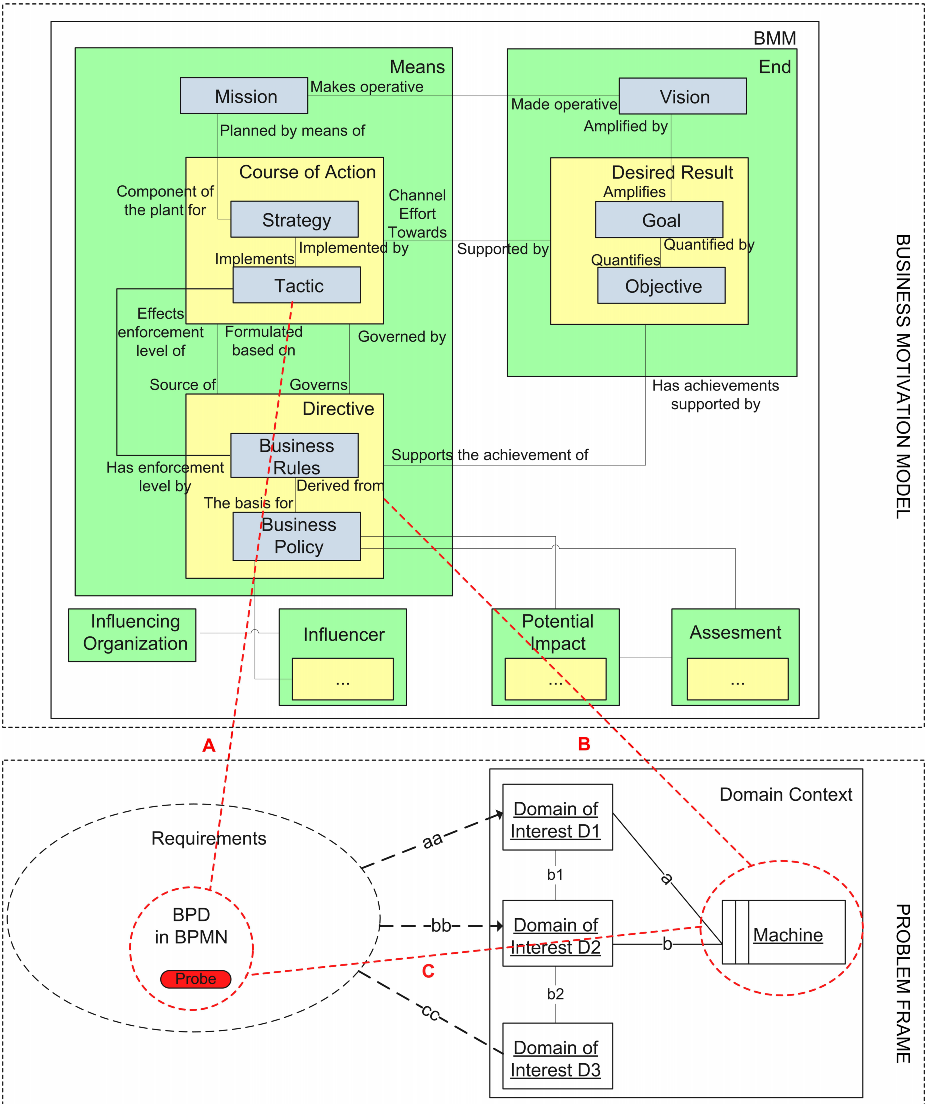


Fig. 1. SAF

3.2 Problem Frames (PF)

Problem Frames (PFs) diagrams provide an analytical framework for requirements based on real world physical entities an their observable interactions and behaviors. PFs were first used to capture, structure and classify recurring software development problems, but they can also be used to define the shape of a business problem.

PF business problem description is done according to two "moods":

- *Indicative* mood represents everything in the problem that is given, and it includes physical domain entities such as people, organizations, departments and devices, and their shared phenomena, such as activities, processes, events, states, commands and information;
- *Optative* mood represents the way we would like every thing to be, it represents the requirements. Requirements include business goals, objectives, processes and all other business and system requirements whose purpose is to alter the 'As Is' view of the world in some way.

Since a requirement can only be understood in the context in which it occurs, a PF diagram consists of two major components: a requirements part and a domain context diagram. Context diagrams contain real world physical domain entities called domains of interest. The phenomena that two or more domains of interest share are indicated by an interface connecting the domains of interest. Shared phenomena consist of observable behavioral phenomena that occur between entities in a context diagram. Context diagrams always contain one special domain of interest, the machine, which is a general purpose computer that is programmed. The requirements part of a problem diagram describes the effects in the real world that the machine should guarantee.

Business Process Modeling Notation (BPMN). An essential component of a business approach strategy is the modeling of activities composing individual business processes. Business process modeling allows the analyst to capture both the broad outline and the specific procedures governing a business. Intuitively, a business process can be seen as the collection of activities that are designed to produce the intended business behavior for a specific customer. The Business Process Modeling Notation [6] is the standard notation used for business process modeling.

Semantics of Business Vocabulary and Business Rules (SBVR). SBVR [20] is a CE formalism proposed by OMG that allows for the formalization of business vocabularies and rules through a combination of syntactical and text formatting rules. SBVR provides a means for describing the structure of the meaning of rules expressed in the natural language that business people use. According to OMG specifications, in SBVR works into logic modalities: *alethic and deontic.*

- SBVR alethic rules are used to model necessities which can not be violated.
- SBVR deontic rules are used to model obligations which ought to be obeyed, but may be violated in real world scenarios.

4 The Comparison with Models in Literature

In this section we make a comparison with other two alignment models, as shown in Table 1, the columns represent the model under comparison while the rows represent which aspect of the comparison has been taken into consideration.

Table 1. Comparison Table

Aspect of Comparison	SAF	GQM+Strategy [2]	PPI Monitoring Model [16]
Strategy	Present Inside the BMM	Present. Limited to the Business Goal, the Software Goal and their related strategies. No Business tattic, vision or objectives is explicated.	None
Measure	None	Present thanks to the GQM approach	None
Directives	Present Inside the BMM	None	None
Probes	Present. Deployed over the BPD, Created to control a Process Strategy, the Logic (including the alarm value) is expressed in the BMM Directives.	None	Present. The structure of the prove is specified in UML. No indication on how to deploy the probe is given.

(Column group heading: Model Compared)

4.1 The Table of Comparison

The first model compared is the GQM$^+$ Strategy Method presented by Basili et al. in [2] and is summarized in section 2.1. The comparison table shows that both GQM$^+$ Strategy and SAF focus on the definition of strategies for business. SAF implements a component namely BMM that empasses the mission, the vision and the strategy of the whole organization, whereas GQM$^+$ Strategy starts from a lower level related to a specific business goal to derive intermediate software goals to be implemented in the traditional GQM. GQM$^+$ Strategy does not implement a control on the process that in SAF is defined by the business directives. As consequence GQM$^+$ Strategy does not implement a control on the business process for its improvement that SAF defines with probes. On the other hand SAF does not provide the user with a rigorous method to derive form the tactic the business directive.

The second model that we call process performance Monitoring Model [16] mainly refer to generate probes for the business process in UML diagrams. In the following section we discuss how the three models can be integrated to provide a more rigorous and complete framework that instantiate business motivation to business process for process performance and improvement in any organization.

4.2 Deriving the Probes in SAF

The novelty introduced in SAF concerns the modeling of monitoring probes over the BPD in accordance with the BMM.

The probes operate over a process, or part of it (e.g. a group of tasks), that has been generated in accordance with a superior business tactic and strategy. The major issue here is to logically configure and allocate the probes in the BP in order to allow the manager monitoring on the achievement of the strategical objectives.

The link between the tactic and the underneath BP gives the position of the probes within the processes, whereas the logical configuration of the probes as control on the BP is supplied by the GQM$^+$ Strategy [2] machine applied to any type of BP - not only software. Specifically, the Business Goal (BG) of the GQM$^+$ Strategy machine is defined in the BMM by the Goal of the Desired Result, while the Measurement Goal (MG) is derived from the Course of Action of the BMM (that is to say from the strategy and the tactic). The rules and the constraints defining the logic of the probes are derived by models relating measures of performance of the BP. The thresholds of performance are then derived with the training of the models on the existing BP and on specific requirements expressed in the tactic. And the logic and the alarm values of the probe reflects the Directives of the BMM in the BP.

In Fig. 1, we show the role of the probes in our framework; the red dotted line named A is used to link a BPD (or part of it) to its generating tactic, this help to focus on a specific part of the business activity. The line indicates "*where do we have to deploy the probe*." With our framework, probes are deployed over the BPD and they do not absolutely affect the topology of it.

The probe deployed in the BP controlls the data flow and eventually reports a warning when a threshold is not met; it generates a report to the linked strategy. The red dotted line named B display the relation between the logic of the probe and the BMM Directives.

The red dotted line named C represents the deployment and the logic link of the Probe within SAF. The PPIMM presented in Section 2.2 by [16] that are modeled in UML can be adopted as model specification for the probes of SAF.

The linking defined by the probes equips SAF of the missing feature of $e^3\!-\!Value$: the relation between the business value analysis and the business strategy.

5 An Application of SAF

In this section we apply SAF to a real case study. A previous validation of the framework has been done in [7] using the EU-Rent case, a fictitious car rental company well known in literature [20]. In particular, in the [7] we showed how to model an organization activity at a fine granularity level.

The case study illustrated below in Subsection 5.1 is based on a real case of an Electric Company. With this case we show the alignment gap that we have found applying our framework in the analysis of both Business Strategy vs. Business Processes and Business Processes vs. Information Technology.

5.1 Electric Company Validation Case

The electric company under study produces and supplies electric power to 118000 users for a total amount of about 1000 GWh per year. Two new European Community directives, 2003/54/CE and 2003/55/CE, has liberalized the electric market since 2007. To comply with the new directives and to remain competitive in this market the electric company has decided to revise its information system in respect of its vision, mission, and business strategy, but under a strong time pressure.

Its vision is to consolidate its image over the territory of a company that takes care of security, ecology and the quality of service. The derived mission is then summarized in three points:

1. Continuously increasing on the quality of the service;
2. Embrace technological innovation;
3. Accurate monitoring of the cost-benefit balance.

The Electric Company finds a disalignment between its mission points and the overall performance of its BPs. We use SAF to surface this gap. A further use of SAF will define and locate appropriate probes into the BPs.

In this study we analyze the BP of the Electric Company as a vendor company that wants to maintain its image in the local market. In this sense, the Electric Company believes that its internal processes that lead to the definition of the invoice represent its core business as a vendor company. Therefore we focus the application of SAF on the *invoice process*.

We start modeling the invoice process with our framework investigating on the possible strategic alignment gap on the invoice controlling procedures. The first step is to model the process activities in respect to the strategic statements. The invoice process is a collaboration (global) BP with four participants, in Figure 3 we focus on the IT Department lane representation that is the most important among the other. The customer receives the invoice notice every two months. In a year, the amount is calculated as follows: the first invoice is calculated on the real kWh consumed in the first bimester; the following two invoices are calculated according to a prediction algorithm; the next invoice is calculated on the real kWh consumed during the two predicted bimester, plus the current bimester, less the kWh already considered in the two previous invoices.

To realize the mission statements above on the invoice process the management has defined the following strategies:

1. Send the invoice notice to the customer regularly (e.g. in the third week of the month) every two months;
2. Produce invoices with ”zero” defects;
3. Provide, in a year, quarterly invoices of almost the same amount.

To implement the strategy in the tactic, the management has supplied the invoice process with a new powerful ERP system. In the execution of the invoice process the mangers still find the following major problem, though:

There is a long delay - of about 45 days - occurring from the time the company collects the real kWh consumed by the customer (with a manual procedure) and the time of the delivery of the invoice. The company wants to understand the causes of this delay and aims to reduce this period that results in a financial loss.

During this context analysis we notice that the company has not defined any performance measurement and regular data collection on the workflow of the invoice process yet. A measurement of performance creates thresholds and rules of performance that allows the management to tune the process during its evolution in time. Only recently a one-shot questionnaire on the task duration has been administrated to the employees. The data of this questionnaire is a starting point and a first overestimation of the overall duration of the invoice process, but reveals a weak of control of the process performance. However, the questionnaire allows us to identify those tasks that mainly affect the duration of the overall process. We find out that on the whole workflow, which is composed of eight activities, seven activities have a mean processing time of about 32 minutes per hundred invoices, the remaining one named *Calcolus Errors and Warning Correction* has a processing time almost six time greater then the mean of the others (See Fig. 2).

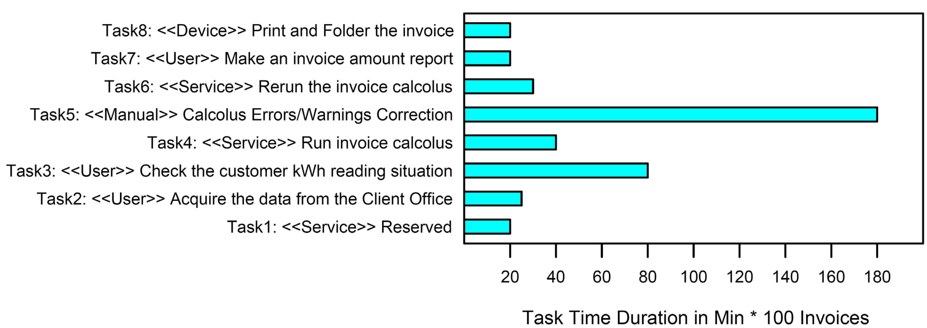


Fig. 2. Invoice Sub-Process Task Time Distribution

From the BPD of the invoice process in Fig. 3 we notice that the activity *Calcolus Errors and Warning Correction* has a good percentage of manual tasks. The main purpose of this task is to discover, check, and correct possible *problems* (also called *warnings*) that derive from the previous tasks that calculate the invoice (for the sequence of the tasks see Fig. 3). These problems are automatically detected by the software system, but the operator responsible for the process must manually check and correct each *warning*. Although the activity to check the process for errors is a crucial activity that expresses directly one

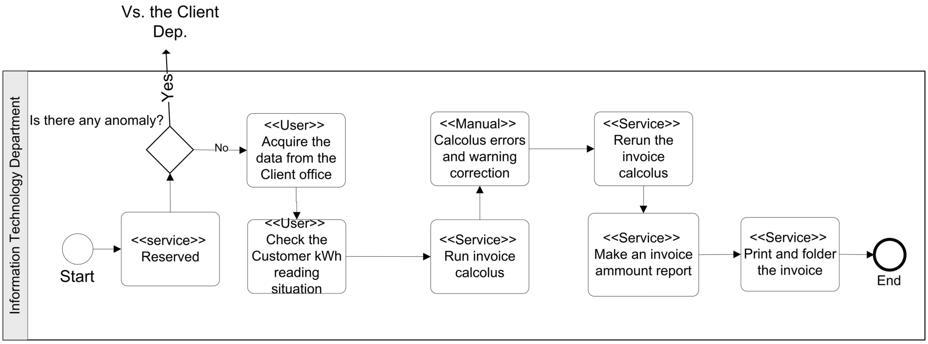


Fig. 3. Invoice Process limited to the Information Technology Department BPMN Lane

of the tactic of the management, it is also the one that consumes the highest percentage of time and it is performed through a checklist filled manually.

This contradiction reveals a misalignment between the management line of control and the operational localization of it in the invoice process. This strategic alignment gap may also be caused by the fact that the current IS is not able to offer support to the tactics defined by the management. The reason can also be found in the recent migration to a powerful ERP system that is not fully used and extended with a workflow engine.

As last point we find that the calculation of task duration in the invoice process is overestimated. We notice that the questionnaire has provided a subjective report of the duration of the invoice process that does not take into account parallel activities and automation of straightforward activities. This analysis has determined problems that have a common root on the non-complete implementation of the strategies in the business process. Neither the presence of a powerful ERP system - that is non-fully exploited though - can solve the gap. A systems of probes to control the performance of the process can monitor and correct the alignment.

6 Conclusions

In this paper, a new strategic alignment framework, SAF has been presented. SAF models business processes in a business context, link them to the organizations' overall goals. SAF also deploys monitoring probes that control BP performance and the achievement of the generating BS.

The framework has been compared with the GQM$^+$ Strategy [3] and Process Performance Indicators Monitoring Model [16] in order to integrate them in SAF. In fact, as described in Subsection 4.2 the GQM$^+$ Strategy is used to define the logical configuration of the probes as control on the BP, while the PPI Monitoring Model can be adopted to describe the design of the probes in UML.

The application of SAF carried out in a case study of an electric company revealed two major alignment gaps. Although the management has supplied the

BPs with a powerful ERP system there is still a misalignment between the management line of control and the operational localization of it in the BPs. This gap may also be caused by the fact that the current IS is not able to offer support to the tactics defined by the management. The recent migration to a powerful ERP system that is not fully used and extended with a workflow engine can be one of reasons. At the time of writing, the authors are applying SAF also on another case study over an italian Local Transportation Company. Form these case studies we aim to further develop and refine SAF supplying evidence of it use.

Acknowledgments

This work was partly funded by the Italian Ministry of Research Fund for Basic Research (FIRB) under project RBNE05FKZ2_004 TEKNE.

We are grateful to P.Eng. Siegfried Tutzer, General Director of Azienda Energetica SpA/Etschwerke AG, and to P.Eng. Emiliano Lutteri, IT Division Head, and his staff for the collaboration offered.

References

1. Basili, V., Caldiera, G., Rombach, D.: Goal question metric paradigm. In: Encyclopedia of Software Engineering, Marcinak, vol. 1, John Wiley and Sons, Chichester (1994)
2. Basili, V., Heidrich, J., Lindvall, M., Münch, J., Regardie, M., Rombach, D., Seaman, C., Trendowicz, A.: Bridging the gap between business strategy and software development. In: ICIS 2007: Proceedings of the International Conference on Information Systems, Montréal, Québec, Canada (2007)
3. Basili, V., Heidrich, J., Lindvall, M., Munch, J., Regardie, M., Trendowicz, A.: Gqm$^+$ strategies – aligning business strategies with software measurement. In: Empirical Software Engineering and Measurement, 2007. ESEM 2007. First International Symposium, September 20-21, 2007, pp. 488–490 (2007)
4. Bleistein, S.J., Cox, K., Verner, J., Phalp, K.: B-scp: A requirements analysis framework for validating strategic alignment of organizational it based on strategy, context, and process. Information Software Technology 48(9), 846–868 (2006)
5. Business motivation model (bmm). Technical report, OMG (2006)
6. Business process modeling notation (bpmn). Misc, OMG (2006), http://www.bpmn.org/Documents/BPMN%20V1-0%20May%203%202004.pdf
7. Ceravolo, P., Fugazza, C., Damiani, E., Mulazzani, F., Russo, B.: Business process monitoring via ontology-based representation models. Technical report (2007)
8. Erdogmus, H., Favaro, J.M., Strigel, W.: Guest editors' introduction: Return on investment. IEEE Software 21(3), 18–22 (2004)
9. Eriksson, H.-E., Penker, M.: Business Modeling With UML: Business Patterns at Work. John Wiley & Sons, Inc., New York, NY, USA (1998)
10. Etien, A., Rolland, C., Salinesi, C.: Measuring the business/system alignment. In: REBNITA 2005: Proceedings of the 1st International Workshop on Requirements Engineering for Business Need and IT Alignment, Paris, France, pp. 143–152. University of New South Wales Press (2005)

11. Gordijn, J., Akkermans, J.M.: Value-based requirements engineering: exploring innovative e-commerce ideas. Requir. Eng. 8(2), 114–134 (2003)
12. Jackson, M.: Problem Frames: Analysing and Structuring Software Development Problems. Addison-Wesley, Reading (2000)
13. Diaz, J.S., De la Vara Gonzalez, J.L.: Business process-driven requirements engineering: a goal-based approach. In: Proceedings of the 8th Workshop on Business Process Modeling, Development, and Support (2007)
14. Kaplan, R.S., Norton, D.P.: The balanced scorecard - measures that drive performance. Harvard Business Review 70(1), 71–79 (1992)
15. Marshall, C.: Enterprise modeling with UML: designing successful software through business analysis. Addison-Wesley Longman Ltd., Essex, UK (2000)
16. Momm, C., Malec, R., Abeck, S.: Towards a model-driven development of monitored processes. In: Oberweis, A., Weinhardt, C., Gimpel, H., Koschmider, A., Pankratius, V., Schnizler, B. (eds.) Wirtschaftsinformatik (2), pp. 319–336. Universitaetsverlag Karlsruhe (2007)
17. Osterwalder, A., Pigneur, Y., Lagha, S.B.: Modeling e-business with ebml. Industrial Organization 0202003, EconWPA (February 2002), http://ideas.repec.org/p/wpa/wuwpio/0202003.html
18. Porter, M.E.: Competitive advantage: creating and sustaining superior performance. Free Press, New York (1985)
19. Reich, B.H., Benbasat, I.: Factors that influence the social dimension of alignment between business and information technology objectives. MIS Quarterly 24(1), 81–113 (2000)
20. Semantics of business vocabulary and business rules (sbvr). Technical report, OMG (2004)
21. Thevenet, L.-H., Salinesi, C.: Aligning is to organization's strategy: The instal-method. In: Krogstie, J., Opdahl, A., Sindre, G. (eds.) CAiSE 2007 and WES 2007. LNCS, vol. 4495, pp. 203–217. Springer, Heidelberg (2007)
22. De Wit, B., Meyer, R.: Strategy - process, content, context: an international perspective. Thomson/South-Western, Mason, Ohio (2004)
23. Yu, E.: Towards modelling and reasoning support for early-phase requirements engineering, pp. 226–235 (1997)

Towards Measuring Key Performance Indicators of Semantic Business Processes

Branimir Wetzstein, Zhilei Ma, and Frank Leymann

Institute of Architecture of Application Systems, University of Stuttgart, Germany
{firstname.lastname}@iaas.uni-stuttgart.de

Abstract. Business Activity Monitoring (BAM) enables continuous, real-time performance measurement of business processes based on key performance indicators (KPI). The performance information is employed by business users but prior support from IT engineers is required for setting up the BAM solution. Semantic Business Process Management (SBPM) tries to minimize the needed support from IT staff throughout the business process lifecycle. In this paper we introduce a framework for BAM as part of SBPM. We show how performance measurement related activities can be integrated into the semantic business process lifecycle. KPIs are modeled by business analysts exploiting semantic annotations of business processes. KPI models are automatically transformed to IT-level event-based models and used for real-time monitoring using reasoning technology.

Keywords: Semantic Business Process Management, Business Activity Monitoring, Key Performance Indicator.

1 Introduction

Business Process Management (BPM) encompasses methods, techniques, and tools that allow organizing, executing, and measuring the processes of an organization. Recently, companies have increasingly been using Business Process Management software to manage their business processes [SF03]. Typically, the business process lifecycle begins with the business analyst analyzing business processes in the company and creating process models using a process modeling tool. When the process is to be automated, it is translated by IT engineers to a workflow model, which is run on a process engine. The process engine executes the workflow model by delegating the process tasks to human workers and automated IT applications.

For controlling the achievement of business goals in business processes and measuring process performance, business activity monitoring (BAM) technology enables continuous, real-time monitoring of processes based on key business metrics, also known as key performance indicators (KPI). Business people define KPIs based on business goals at process design time. These KPIs are then translated to IT monitor models by IT engineers. At process execution time, event-based BAM technology displays information on KPIs in dashboards and alerts responsible business managers in case of severe deviations from planned values.

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 227–238, 2008.

Despite of increasing software support for BPM in general and BAM as part of BPM in particular, there is still a low degree of automation in the business process lifecycle [HLD+05]. In particular, there are substantial difficulties when it comes to bridging the gap between the business and IT views of business processes. One of the major problems is the translation of the high-level business process models and corresponding KPI models, which are modeled by a business analyst, to workflow models and monitor models on the IT level. The vision of Semantic Business Process Management (SBPM) is to close the Business-IT gap by using semantic technologies [HLD+05].

In this paper we introduce a framework for modeling and monitoring of KPIs in SBPM. Our goal is to increase the degree of automation in measuring of KPIs by enabling the business analyst to define the KPIs and then generate the IT artifacts needed for monitoring. The contribution of the paper is as follows. (i) We integrate the KPI management into the overall semantic business process (SBP) lifecycle by depicting how KPIs are modeled, implemented and monitored in the respective SBP phases. (ii) We create a KPI ontology which enables the business analyst to define the KPIs based on ontology concepts which are part of semantic annotations of semantic business processes. Semantic business processes explicitly specify the semantics of process activities by modeling their inputs, outputs, preconditions and postconditions, in terms of business objects (e.g., “purchase order”) and their state changes (e.g., “purchase order received”). As KPIs are also based on business objects (e.g., “percentage of purchase orders which were delivered successfully and on time”), semantic annotations of processes can be exploited for the definition of KPIs in the SBP modeling phase. (iii) We depict how the KPI model is transformed to a monitor model, which supports evaluation of KPIs based on events published by the process engine at process execution time. As the KPIs are defined on an ontological level, we show how reasoning technology is used for their computation.

The rest of the paper is organized as follows. In Section 2 we depict how our performance management framework fits into the overall SBP lifecycle. Section 3 describes how KPI models are created in the SBP modeling phase. Section 4 describes how KPI models are transformed to executable monitor models and how they are measured at process execution time. In Section 5 we present related work, and finally, in Section 6 we outline our future work.

2 Performance Management within the SBP Lifecycle

In [Fa07] the SBP lifecycle has been defined as consisting of four phases: SBP modeling, SBP configuration, SBP execution, and SBP analysis. Apart from these four phases of the SBP lifecycle, there is a business strategy definition layer which is providing the basis for SBPM by defining the business strategy and business goals for accomplishing that strategy. Business goals are broken down to key performance indicators which are key business metrics for measuring the business performance. By measuring the KPIs, responsible business people control whether the business goals are achieved as planned.

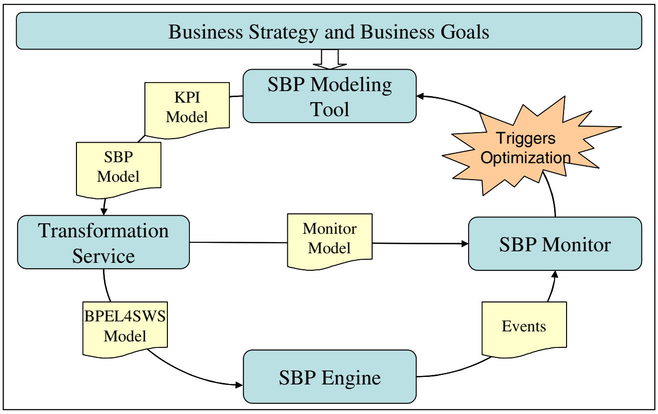


Fig. 1. Performance Management Lifecycle

In the first phase of the SBP lifecycle, the SBP modeling phase, business analysts model business processes and corresponding KPIs based on the requirements posed by the business goals in the strategy layer. [1] Business analysts use the SBP modeling tool to create a process model based on a process modeling notation such as BPMN [BPMN06] and annotate it semantically by referencing concepts from domain ontologies. These concepts describe business objects which are worked on in the business processes. The business analyst specifies the KPIs based on the business objects and their state changes. A KPI is related to the process model by referencing process activities and their semantic annotations. A KPI contains a computation definition and a target value which is to be accomplished to satisfy a business goal. In addition, allowed ranges of values and in case of deviations the resulting notifications can be defined. The resulting artifacts in the SBP modeling phase are an SBP model and a corresponding KPI model.

In the SBP configuration phase the SBP model is transformed to an executable workflow model, in our case BPEL4SWS [NLK+07], which is an extension of BPEL towards interaction with Semantic Web Services. The KPI model is automatically transformed to a monitor model which is an IT-level representation of the KPI model. In addition, deployment information for the SBP engine is generated. It specifies for which activities of the BPEL4SWS process model logging events should be published at process runtime. The SBP monitor subscribes to these events and evaluates the KPIs based on them at runtime. The executable BPEL4SWS process model is deployed to the SBP engine; the monitor model is deployed to the SBP monitor.

In the SBP execution phase, the SBP engine executes the process and publishes events to a publish/subscribe event infrastructure. The SBP monitor receives the events and computes the corresponding KPI values. As the KPIs are based on semantic descriptions, a reasoner is used for their evaluation. In case of deviations from allowed values, notifications are sent to responsible business users. The KPI values are displayed in dashboards.

[1] The relationship between business goals at the strategy layer and KPIs specified at the process modeling layer is out of scope of this paper.

Finally, in the SBP analysis phase, the KPI values are analyzed and the achievement of business goals is evaluated. In case of non-satisfying performance, causes are tried to be found, which eventually can lead to optimization of processes in the SBP modeling phase, thus closing the SBP lifecycle. The SBP analysis phase is out of scope of this paper.

3 Modeling Key Performance Indicators

KPIs are key business metrics for getting information on the performance of a company. In this paper we concentrate solely on KPIs which evaluate the efficiency and effectiveness of business processes and which can directly be derived and computed based on the runtime data of the executable business processes. Other KPIs, e.g. financial metrics such as Return on Investment (ROI) are not directly measurable by using runtime data of business processes and thus are out of the scope of this paper.

A KPI can be seen as a key process metric, which is evaluated in a certain analysis period and has a target value which should be reached or preserved within the analysis period indicating the achievement of predefined business goals [SS06, WAM+07]. Typically, in addition to a target value, allowed ranges of values are defined. In case of deviations, notifications are sent to business people in real-time. An example for a KPI is “percentage of purchase orders that were processed in time in the last 3 months should be greater than 90%”, the analysis period being 3 months and the target value 90%.

Our analysis of concrete KPIs, such as SCOR Supply Chain Metrics [SCOR07], and examples in related work [MR00, SS06], shows that most KPI computations are specified based on business objects which are dealt with in the process to be measured. Such business objects (a.k.a. business items, business documents or business data) are for example “purchase order”, “invoice”, or “shipment”. Typically, KPIs contain a condition on the state of these business objects or query their attributes, e.g. “number of purchase orders processed successfully and in time”.

Figure 2 shows two typical KPI templates, which are based on business objects [SS06]. For example, for a “purchase order process” the “business object” would be replaced by “purchase order”.

In order to be able to model KPIs for business processes, one has to know which business objects are handled by a process and which state changes occur at which

KPI	Computation Expression
Deadline Adherence	$\frac{\text{Number of finished Business Objects on time}}{\text{Number of all finished Business Objects}} \times 100\%$
Process Duration	$\frac{\sum \text{(finish date - start date) of all finished Business Objects}}{\text{Number of all finished Business Objects}}$

Fig. 2. KPI Templates based on Business Objects

activities. Often, however, business objects and their state changes are not explicitly modeled as part of business process models in the process modeling phase. In most cases they are implicitly part of the names of activities, e.g. “Receive Purchase Order”. In semantic business processes, however, activities specify the semantics of their inputs, outputs, preconditions and postconditions[2][3] explicitly by referencing concepts defined in domain ontologies. These concepts model business objects and information on their state changes [BDW07]. Semantic business processes thus explicitly model at which activities business objects are processed and how their states change. Therefore, as the information needed for KPI specification is provided already in the process modeling phase, we can pursue a model-driven approach: the KPIs are specified by business analysts as a mathematical formula based on business objects as sketched in Figure 2 and then automatically transformed to an event-based monitor model which is used for their evaluation at process execution time. When it comes to tool support for the business analyst, besides providing support for entering a mathematical formula, the tool can present the business objects and the possible state changes which are part of semantic annotations of process activities automatically and thus help the analyst determine the objects and states he can use in the formula. Implementing support for specifying KPIs as part of an SBP modeling tool is part of our future work.

3.1 KPI Ontology

For the specification of KPIs we have defined a KPI ontology using WSML [BLK+05] as the ontology language.[4] Figure 3 shows the KPI ontology and its relation to the relevant concepts from the process modeling ontology [He07] and the event ontology [PD07]. A KPI model groups KPIs for a business process. A KPI consists of an aggregate metric definition, which computes the actual value of the KPI in a certain analysis period. The target value of the KPI specifies a value which is to be achieved in the analysis period. For deviations from target values alerts can be defined.

When formulating metrics for KPIs we can distinguish between instance metrics and aggregate metrics. An instance metric calculates metrics for one single process instance of a process model. Instance metrics measure duration between two activities in the process or test a condition on the state of the process instance at a certain point in the process. The point in time when instance metrics are to be evaluated is modeled by execution events. Execution events are part of the event ontology and provide the needed link of the KPI computation to the process model. At process runtime these events are published by the process engine before and/or after the execution of

[2] Pre/postconditions denote the “state of the world” before and after the execution of the activity. In particular, a pre/postcondition can contain assertions which are not reflected in the input/output of the activity, i.e. as an attribute of a business object.

[3] We have chosen technology-neutral terms here. In WSMO, the corresponding terms are: preconditions, postconditions, assumptions and effects.

[4] When evaluating KPIs, we will not reason over the KPI ontology, but over domain ontologies which contain the business objects. Nevertheless, we have used an ontology language for the KPI metamodel because it is interlinked with the process ontology and event ontology, and all SBPM-related metamodels in our framework are defined as WSML ontologies [He07].

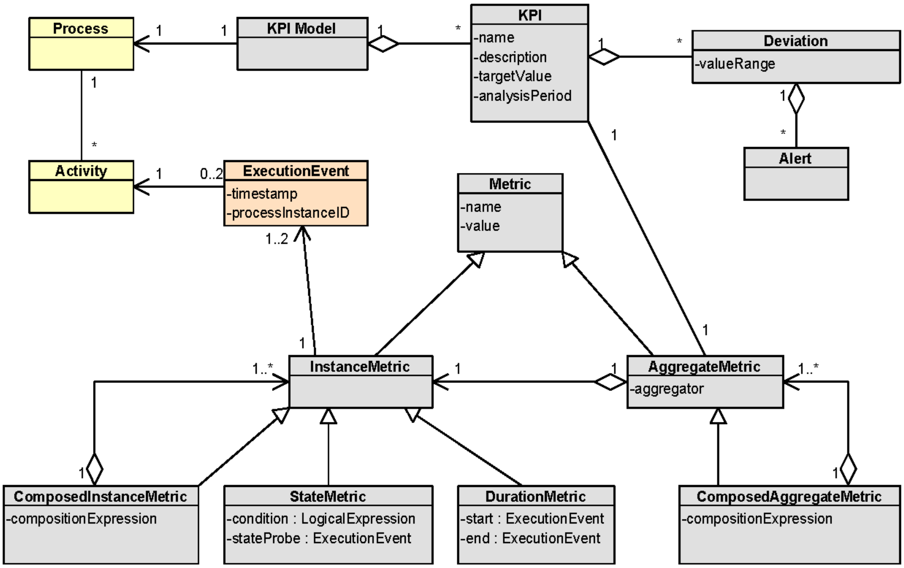


Fig. 3. KPI Ontology as a UML Class Diagram (simplified)

process activities and are received by the process monitoring tool for calculation of the KPIs. The execution events provide the attribute “timestamp” which enables the calculation of the duration between activities. The duration metric measures the time between two activities by simply subtracting the “timestamp” attribute values of two execution events.

A state metric references an execution event and specifies a logical expression which formulates a condition on the state of the process instance at that point in time. The state of the process instance at a certain point in time is given on the one hand by the values of the business objects, and on the other hand by assertions which are part of the postconditions but are not directly modeled as attributes of business objects; e.g. we could define a WSML relation “OrderDeliveredSuccessfully(Order)” which acts as an assertion in the postcondition of an activity to express successful processing of a purchase order process. The state metric condition evaluates to true or false. The logical expression is formulated in the ontology language which is used for the specification of the domain ontologies, e.g. in WSML, and constitutes thus an “intelligent query” of the state of the process instance. In addition to specifying duration and a state condition, instance metrics can be composed using arithmetic and logical operators. Thus, a duration metric could alternatively be defined as a subtraction of the execution event timestamps of two state metrics; we have defined an explicit duration metric concept for convenience reasons.

Aggregate metrics aggregate values of instance metrics using operators “max”, “min”, “average”, and “sum”. They thus evaluate the performance of a process in a certain analysis period. For example, if a state metric evaluates whether a “purchase order has been processed in time”, an aggregate metric would evaluate “the number (sum) of purchase orders processed in time” in a certain analysis period. Aggregate metrics can be composed using arithmetic operators.

An important aspect of KPI monitoring is the notification of business users in case of deviations from target values. We define deviations by specifying numerical ranges around the target value which the actual value is allowed to be in. In case of a deviation from this range of values an alert is raised, e.g., by sending an email, invoking a web service, or displaying a message in the dashboard.

3.2 Example

In the following we will demonstrate the use of the KPI ontology based on an example business process. Figure 4 shows a business process for processing purchase orders: After receiving a purchase order, the order is first analyzed. If it can be processed without any changes an acknowledgement is sent, otherwise a new offer is sent to the customer, resulting in a changed purchase order. In the next step the purchase order is processed, shipped, and finally a shipment receipt acknowledgement is received from the customer.

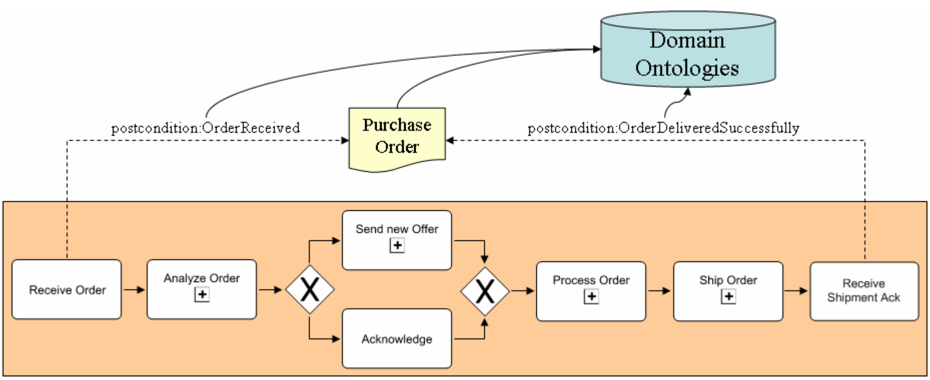


Fig. 4. Purchase Order Process

We define two KPIs which measure deadline adherence and process duration for that process (Figure 5). For the definition of the two KPIs we make following assumptions: the purchase ontology defines an “Order” concept, which contains an attribute “assuredDeliveryDate”; the purchase ontology defines (WSML) relations for the different states of the “Order” concept, e.g. “OrderReceived(Order)” and “Order DeliveredSuccessfully(Order)”; these relations are used in postconditions of process activities, e.g. after the last activity either the relation “OrderDelivered Successfully(Order)” or “OrderDeliveredLate(Order)” could hold; the purchase ontology contains additional relations which categorize orders based on products they contain, e.g. “ElectronicProductOrder(Order)” for orders that contain only electronic products.

For calculating the process duration we measure the time between the end and start activity by specifying a duration metric referencing the two corresponding execution events. An aggregate metric calculates then the KPI value by aggregating the duration metric values using the “average” operator.

We measure deadline adherence as the percentage of orders delivered successfully and on time whereby we are interested in orders that only contain electronic product items. The computation formula of that KPI (Figure 5) is mapped to five metrics: an instance state metric to evaluate whether a purchase order is processed successfully and on time, a corresponding aggregate metric which sums up the instance metrics which evaluate to “true”, an instance state metric which evaluates to “true” when a new purchase order is received, the corresponding aggregate metric which counts those occurrences, and finally, an aggregate metric which composes the two former aggregate metrics to compute the percentage value.

KPI	Computation Expression
Deadline Adherence	$\dfrac{\sum \begin{array}{l}\text{ShipmentAckEvent.timestamp} <= \text{Order.assuredDeliveryDate} \\ \text{AND OrderDeliveredSuccessfully(Order)} \\ \text{AND ElectronicProductOrder(Order)}\end{array}}{\sum \text{OrderReceived(Order)}} \times 100\%$
Process Duration	Avg (ShipmentAckEvent.timestamp – ReceiveOrderEvent.timestamp)

Fig. 5. KPI definitions for the Purchase Order Process

The benefit of using ontologies in our approach is shown when evaluating state metrics. In order to define a state metric which evaluates whether the purchase order for electronic products has been processed successfully and on time, we need to test a condition on the last activity of the process. We therefore define the execution event “ShipmentAckEvent” which occurs after the activity “Receive Shipment Acknowledgment” has been processed. For testing whether the purchase order was processed on time, we compare the “assuredDeliveryDate” with the timestamp of the execution event. For testing whether the purchase order has been processed successfully, we check whether the relations “OrderDeliveredSuccessfully(Order)” and “ElectronicProductOrder(Order)” hold. The WSML Logical Expression for the instance metric looks as follows (namespaces omitted):

```
?order[assuredDeliveryDate hasValue ?plannedDate] memberOf Order
    and ?event[activity hasValue ?activity] memberOf ExecutionEvent
    and ?activity[name hasValue “Receive Shipment Acknowledgement”]
    and ?event[timestamp hasValue ?eventDate]
    and (dateLessThan(?eventDate, ?plannedDate)
        or (dateEqual(?eventDate, ?plannedDate))
    and OrderDeliveredSuccessfully(?order)
    and ElectronicProductOrder(?order)
```

The WSML logical expression is automatically generated based on a formula, which the business analyst specifies with tool support (as sketched in Figure 5). It is then utilized as a query to the WSML reasoner at process runtime (Section 4). The benefit of our approach is that implicit knowledge can be taken into account when evaluating state metrics. In our example, based on the WSML relations, which are defined by appropriate WSML axioms, the reasoner would infer whether the order is an electronic product order based on a product ontology and evaluate the postcondition of the activity.

4 Monitoring Key Performance Indicators

After the semantic business process and KPIs have been modeled in the SBP modeling phase, they are transformed to executable artifacts in the SBP configuration phase. This involves the transformation of the SBP model to an executable BPEL4SWS process model, and the transformation of the KPI model to a monitor model for the SBP monitor and SBP engine specific deployment information on which events have to be published at runtime.

When KPI models are specified for an SBP model, in addition to the generation of the executable BPEL4SWS process model, information is needed for specifying how the execution events which are bound to activities and are used for the evaluation of instance metrics are published at process execution time. There are basically two approaches how to generate these events: (i) The generated BPEL4SWS process model does not contain any event specific logic, i.e. events are transparent to the BPEL4SWS process model. A BPEL engine typically already provides a mechanism for logging the execution of process instances by using events based on an internal event model. These events are stored in an audit log. Additionally, events can be published on a publish/subscribe infrastructure to enable real-time monitoring, which is what is required in our case. As only a small fraction of possible events is needed for KPI evaluation, deployment information for the SBP engine is generated which specifies for which activities events are to be published at runtime. This information is derived from the execution event definitions in the KPI model; the deployment information is specific to the SBP engine implementation, as the BPEL specification does not specify an event model. In our case the events are execution events defined in the event ontology and contain instance data of business objects which are defined in domain ontologies. The SBP engine is adapted to generate this kind of events. (ii) The generated BPEL4SWS process model contains explicit activities which publish events to a predefined interface provided by the SBP monitor. The events contain the specific data which is needed for the computation. The event data cannot however contain technical information such as the time of the activity execution. Also the BPEL4SWS process now contains much technical information which makes it less readable. Therefore, we have chosen the first approach.

The KPI model is transformed to a monitor model. The monitor model specifies how events received by the SBP monitor are processed to calculate the KPIs. For event handling, complex event processing (CEP) technology could be used [Lu02]; the implementation specifics are part of our future work. In the SBP execution phase, the SBP engine publishes events which are received and evaluated by the SBP monitor to calculate the KPI values and display them in dashboards (Figure 6). Instance metrics are evaluated based on incoming events, which the SBP Monitor subscribes to. An event contains the identifier of the corresponding process instance.

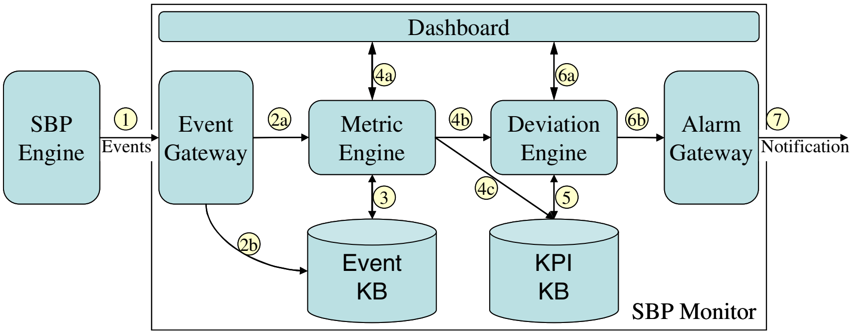


Fig. 6. SBP Monitor Architecture

When the process monitor receives an event, the event and corresponding data (business objects) that the event contains, are inserted into the knowledge base (KB) of the corresponding process instance. The metric engine determines in which instance metrics the event is used, as more than one instance metric could use this event. To evaluate a state metric, the condition of the state metric which is formulated as a logical expression acts as a query to the reasoner. In our example, the reasoner calculates whether the order was delivered on time, checks whether "OrderDelivered-Successfully(?order)" holds, and infers whether all items of the order are electronic products. For the evaluation of a duration metric, it has to be waited until the second event with the end date is received, until the computation can be performed. After the evaluation of the instance metrics all other dependent metrics are evaluated, i.e. instance metrics are evaluated first; followed by composite instance metrics; aggregate metrics are updated every time a new aggregated instance metric has been calculated; finally, composite aggregate metrics are computed. When the corresponding aggregate metric has been computed, the value of the KPI is updated in the KPI KB. After every update of the KPI's actual value, the deviation rules are checked. In case of a deviation from the allowed value ranges, the deviation engine sends a notification to a business user (e.g. an email), shows a message in the dashboard, or invokes a web service.

5 Related Work

There are several approaches to performance management based on KPIs and BAM described in the literature and already implemented in products. IBM's approach integrates performance management tightly into the BP lifecycle in a similar manner as we do, and supports it through its WebSphere family of products [WAM+07] in a model-driven way [CJ06]. However, the computation formula has to be specified by IT developers in the workflow layer. In our approach the KPI is completely defined by the business analyst based on the semantic annotation of the business process model. In [CCD+04], the HP Business Process Cockpit product is described which supports KPI definition by business analysts based on so called behavior templates (similar to the KPI templates shown in Figure 2). However, behavior templates have to be predefined by IT engineers using SQL queries over the audit trail. None of these approaches uses ontologies for the description of business objects, and thus does not support inferring of implicit knowledge during metric evaluation.

In [APA+07] the vision of semantic process mining and monitoring is outlined. Semantic process mining deals with analysis of already executed process instances; process monitoring analyzes executing process instances. For SBP Monitoring a 5-phases "Observe-Evaluate-Detect-Diagnose-Resolve" approach is suggested. In this paper we deal with the first three phases: observation of events, evaluation of KPIs, and detection of deviations and resulting alerts.

The BAM approach described in this paper should be differentiated from other forms of business process analysis. BAM denotes real-time event-based monitoring, usage of dashboards to display business-level metrics to business decision makers, and automatic notification in case of deviations. This kind of monitoring is also called "active monitoring" as opposed to "passive monitoring" [MR00]. In passive monitoring the events and metrics are not actively evaluated by the system, but are explicitly issued by the user. The user can do this e.g. by ad-hoc querying the audit trail or asking the process engine about information on the running process instances. Completely other analysis techniques are process mining and data mining [CCD+04] based approaches. These techniques go beyond just reporting what is happening in the processes, but try to find out the causes and predict future behavior.

6 Conclusions and Outlook

In this paper we have presented an approach to semantic business process performance management based on key performance indicators.[5] Compared to existing work, the benefit of our approach is (i) that KPIs can be fully specified by business analysts in the process modeling phase, because the needed information, namely business objects and their state changes, is available as semantic annotations of activities; (ii) as KPIs are specified based on business objects which are defined in ontologies, machine reasoning can exploit implicit knowledge during their evaluation. While the former aspect could also be implemented in non-semantic approaches, because it only requires that business objects (not necessarily defined in ontologies) are part of annotations of process models, the latter aspect is constrained to the use of ontologies.

In our future work we will implement the described approach by extending an existing SBP modeling tool and building an SBP monitor. In addition we will extend the KPI ontology by adding the possibility to specify process cost based KPIs and KPIs which are based on more than one business process. Apart from real-time active monitoring, we will investigate how semantics can be used for ad-hoc queries over audit trails (passive monitoring).

---

[5] This work has in part been funded through the European Union's 6th Framework Programme, within Information Society Technologies (IST) priority under the SUPER project (FP6-026850, http://www.ip-super.org).

References

[APA+07] Alves de Medeiros, A.K., Pedrinaci, C., Aalst, W.M.P., Domingue, J., Song, M., Rozinat, A., Norton, B., Cabral, L.: An Outlook on Semantic Business Process Mining and Monitoring. In: Proceedings of 3rd International IFIP Workshop On Semantic Web & Web Semantics (SWWS 2007) at On The Move Federated Conferences and Workshops (2007)

[BDW07] Born, M., Doerr, F., Weber, I.: User-friendly Semantic Annotation in Business Process Modeling. In: Benatallah, B., Casati, F., Georgakopoulos, D., Bartolini, C., Sadiq, W., Godart, C. (eds.) WISE 2007. LNCS, vol. 4831, Springer, Heidelberg (2007)

[BLK+05] de Bruijn, J., Lausen, H., Krummenacher, R., Polleres, A., Predoiu, L., Kifer, M., Fensel, D.: The Web Service Modeling Language WSML (October 5, 2005), http://www.wsmo.org/TR/d16/d16.1/v0.21/

[BPMN06] Business Process Modeling Notation Specification. OMG Final Adopted Specification, February 6 (2006)

[CCD+04] Castellanos, M., Casati, F., Dayal, U., Shan, M.-C.: A Comprehensive and Automated Approach to Intelligent Business Processes Execution Analysis. Distributed and Parallel Databases Journal 16(3) (2004)

[CJ06] Chowdhary, P., Jeng, J.J., et al.: Model Driven Development for Business Performance Management. IBM System Journal 45(3) (2006)

[Fa07] Fantini, P. (ed.): Semantic Business Process Lifecycle. SUPER Project Deliverable D2.2 (May 2007), http://www.ip-super.org/res/Deliverables/M12/D2.2.pdf

[He07] Heymans, S. (ed.): Business Process Ontology Framework. SUPER Project Deliverable D1.1 (May 2007), http://www.ip-super.org/res/Deliverables/M12/D1.1.pdf

[HLD+05] Hepp, M., Leymann, F., Domingue, J., Wahler, A., Fensel, D.: Semantic Business Process Management: A Vision Towards Using Semantic Web Services for Business Process Management. In: Proceedings of the IEEE ICEBE 2005, Beijing, China, October 18-20, 2005, pp. 535–540 (2005)

[Lu02] Luckham, D.C.: The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (2002)

[MR00] zur Muehlen, M., Rosemann, M.: Workflow-based Process Monitoring and Controlling - Technical and Organizational Issues. In: HICSS 2000 (2000)

[NLK+07] Nitzsche, J., van Lessen, T., Karastoyanova, D., Leymann, F.: BPEL for Semantic Web Services. In: Proceedings of the 3rd International Workshop on Agents and Web Services in Distributed Environments (AWeSome 2007) (2007)

[PD07] Pedrinaci, C., Domingue, J.: Towards an Ontology for Process Monitoring and Mining. In: Semantic Business Process and Product Lifecycle Management (SBPM 2007), Innsbruck, Austria (June 2007)

[SCOR07] Supply Chain Operations Reference Model, SCOR Version 8.0 (2007), http://www.supply-chain.org

[SF03] Smith, H., Fingar, P.: Business Process Management. The Third Wave. Meghan-Kiffer, US (2003)

[SS06] Schmelzer, H.J., Sesselmann, W.: Geschäftsprozessmanagement in der Praxis. Hanser Verlag (2006)

[WAM+07] Wahli, U., Avula, V., Macleod, H., Saeed, M., Vinther, A.: Business Process Management: Modeling through Monitoring Using WebSphere V6.0.2 Products. IBM Redbook SG24714801 (2007)

Structural Detection of Deadlocks in Business Process Models

Ahmed Awad and Frank Puhlmann

Business Process Technology Group
Hasso Plattner Institut
University of Potsdam, Germany
{ahmed.awad,frank.puhlmann}@hpi.uni-potsdam.de

Abstract. A common task in business process modelling is the verification of process models regarding syntactical and structural errors. While the former might be checked with low efforts, the latter usually requires a complex state-space analysis to prove properties like deadlock-freedom of the models. In this paper, we address the issue of deadlock detection with a novel approach based on business process querying. Using queries, we are able to detect a broad range of common structural errors that lead to deadlocks, such as misaligned splits and joins. While not being complete, the proposed approach has the advantages of low computational efforts as well as providing graphical outputs that directly lead to the errors.

1 Introduction

With the maturing of business process management (BPM) as an integrated approach ranging from modeling over execution up to evaluation [3], the verification of process models becomes ever more important. This is obvious, since the direct execution of process models requires the detection of errors before the process models are executed. Typical errors can be classified into two categories, either syntactical or structural. A syntactical error is given if modeling elements are used in an invalid manner. The valid and invalid combinations are usually prescribed by the corresponding standard that is used. The Business Process Modeling Notation (BPMN), for instance, does not allow events with more than one outgoing arc [1]. Syntactical errors can usually be found within reasonable time by simply parsing through the process model. Structural errors, such as given by misaligned splits and joins, cannot be detected easily, since the runtime behavior of the process needs to be investigated. To achieve a complete judgement whether a process model fulfills certain structural criteria—such as deadlock freedom—usually the complete state-space has to be analyzed. This analysis, however, is costly in terms of required memory and computing time. In some cases, the result cannot be computed at all [2].

In this paper, we address the structural analysis of business process models regarding deadlocks. A deadlock in a process model is given if a certain instance of the model (but not necessarily all) cannot continue working, while it has

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 239–250, 2008.

not yet reached its end. The proposed approach is based on graphical queries given in BPMN-Q as introduced in [7]. A BPMN-Q query is represented a as small business process diagram that might contain additional query elements that will be substituted with BPMN elements during its processing. The result of such a graphical query is given by a sub-graph of the original process model. We discuss the idea of detecting deadlocks in process models by formulating—and evaluating—queries that only result in non-empty sub-graphs of the queried process models if a deadlock is contained. For this purpose, we present a set of so called *deadlock patterns* whose occurrence in process models usually leads to deadlocks.

While the approach is not complete—in a sense of finding all possible deadlock sources—each matching query relates to a structural error in the process model. The approach has two major advantages. First, we assume that is computable in polynomial time, meaning that a majority of structural errors are actually detectable. A formal proof, however, is ongoing research. Second, if an error is found, it provides a direct graphical output leading to the error—the resulting sub-graph of the query. In contrast to state-space analysis that requires high efforts and is sometimes not computable, our approach is actually suited to support business process modelers in finding errors in their process models.

The paper is structured as follows. We first extend the motivation and discuss related work in section 2. In section 3 the preliminaries—that is BPMN-Q, the graphical query language, as well as existing deadlock pattern—are introduced. The contribution is presented in section 4, where we discuss BPMN-Q-based deadlock queries. Section 5 gives a larger example. Finally, the paper is concluded with a discussion of future work in section 6.

2 Motivation and Related Work

As already stated, the major motivation of our work is given by lowering the required efforts for detecting deadlocks in process models. Obviously, a full state-space analysis in a transition system underlying a graphical process model (e.g. Petri nets, π-calculus) is sufficient and complete. The drawbacks—in many cases—are the high computational efforts as well as the typically binary result. The backtracking of transition sequences to actual elements of a graphical process model is a complex process in itself. Our approach, in contrast, focuses on the graphical model, allowing the designer to receive direct feedback on erroneous part(s) of the process model. The technique that we apply is based on graphical queries denoted in BPMN-Q. While BPMN-Q in itself is very helpful for process modelers, e.g. in searching existing process models, we discuss how BPMN-Q can be applied to detect deadlocks in process models. Hence, we also provide an additional application area for graphical process queries.

Our approach, however, does not resemble existing soundness properties [2]. We are not able to cover even weaker variants—such as lazy or weak soundness [8,6]—since we do not have the possibilities to query all quantifications based on structural properties only.

Analyzing the structure of a business process to detect deadlocks has already been addressed in literature. In [9], the authors introduced a set of reduction rules for the process graph as means to indicate the correctness. When a process graph can be reduced to an empty graph, the process is said to be conflict free. On the other hand, if the reduction algorithm fails to reduce the process graph to an empty graph, this means the presence of a conflict. One limitation of this approach is that it works on acyclic process graphs only. Another limitation is, that when an identification of a conflict occurs, the reduced graph possibly looks different from the original graph (due to elimination of nodes and edges) which might make the visual identification of the problem difficult. Another approach called causal footprints was introduced recently in [10]. It is able to detect deadlock, trap, and multiple termination patterns by mapping the structure of Event-driven Process Chains [5] into the notion of these causal footprints. Afterwards, reasoning about properties of the resulting causality graphs is possible. Both mentioned approaches, however, have the drawbacks that all detection rules are hard-coded into the supporting tools. This means that any attempt or discovery of further patterns necessitates the modification of source code in these tools.

3 Preliminaries

This section introduces the preliminaries required for detecting deadlocks using BPMN-Q. In particular, BPMN-Q is introduced, existing deadlock patterns are discussed, and limitations and assumptions are given.

3.1 BPMN-Q

BPMN-Q [4] is a visual language that is based on BPMN [1]. It is used to query business process models based on their structure. Beside the set of notations defined in BPMN, BPMN-Q extends them with seven new elements. Some of these elements are flow objects, the others are for connectivity. These elements are shown in figure 1 and are described as follows:

(a) *Variable Node*: it resembles an activity but is distinguished by the @ sign in the beginning of the label. It is used to indicate unknown activities in a query.
(b) *Generic*: this indicates an unknown node in a process. It could evaluate to anything—even null—except for start events.
(c) *Generic Split*: a generalization of any type of split gateway.

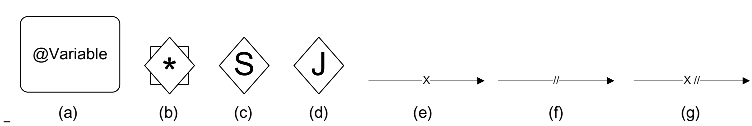


Fig. 1. Extended Elements of BPMN-Q

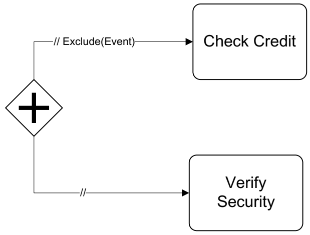


Fig. 2. Sample Query

(d) *Generic Join*: a generalization of any type of join gateway.
(e) *Negative Sequence Flow*: states that two nodes A and B are not directly related by sequence flow.
(f) *Path*: states that there must be a path from A to B. A query usually returns all paths.
(g) *Negative Path*: states that there is not any path between two nodes A and B.

An example is shown in figure 2. It shows a simple query that detects if a given process model contains the tasks *Check Credit* and *Verify Security* in a parallel order. Since arbitrary nodes might be in-between the AND-gateway and the required tasks, path query elements are used. A path might have an additional path expression, written in brackets on the path. The only path expression that we use in this paper is the exclusion of certain nodes from a path, denoted as *exclude(NodeType)*. The upper path of the example excludes intermediate events from the path for illustration purposes.

3.2 Deadlock Patterns

Deadlock patterns have already been identified by Onada et al. in [7]. Two concepts were behind these patterns. The first is *reachability*. Reachability between two nodes A and B in a process graph simply means that there is at least one path from A to B. The second is *absolute transferability*. This is a much stronger concept because it states that a token (work item) can always be transferred from node A to *all* input points of node B. What makes absolute transferability reduce reachability between two nodes is the existence of *routing* control nodes in between. From the semantics of AND-join nodes, we know that a deadlock occurs if not all its input points are activated. If we analyze the type of connectivity between an AND-join node and its sources with respect to the absolute transferability and reachability concepts, we can come out with three combinations:

– *Reachability with absolute transferability*. The execution path from a node A to the input points of an AND-join is free of XOR and OR splits. According to the definition of absolute transferability, there is no chance for deadlocks.
– *Reachability without absolute transferability*. Here the execution path from some node A to the input points of an AND-join node includes XOR (OR) splits. Here, there is a possibility for deadlocks.

– *No Reachability*. This means that there is no path from a certain node to the inputs of an AND-join, so no chance for deadlocks to occur.

It is now clear, that whenever there is a reachability without absolute transferability, there is a chance for a deadlock.

The authors of [7] also identified the possibility for a reachability without absolute transferability from the output of an AND-join back to its input. They called this behavior the *loop deadlock type* (see Table 3 row 4 in [7]). Another pattern they identified is the *multiple source deadlock type*. It occurs when an AND-join has input that stems from different sources (see Table 3 row 5 in [7]). According to the above discussion, five deadlock patterns were identified. We represent them next (in a compact form as three of them are variants of each other) along with a discussion when applying them to BPMN models.

– *Loop*: occurs when there is an execution path from the output of an AND-join back to a subset of its input points. If this path contains an XOR-split, deadlock occurs only when the branch leading to the loop is chosen. In case there is a path that does not contain XOR-splits deadlock occurrence is certain.
– *Multiple Source*: occurs when an AND-join has input points which are at some point in the process up-stream originate from two different sources. Assuming that none of the source nodes is the AND-Join itself, we can see that the multiple source pattern can occur (distinctly from other pattern) only when the process structure is one of the following:
  - One of the two sources is an XOR-split. This specification intersects with the third type of patterns shown below.
  - The process has multiple start points that are later on synchronized. In case of models specified in BPMN, multiple starts are permissible. Actually, multiple start points resemble an AND-split between the start events, hence we can deduce that there is reachability between two or more sources (start events) to the AND-join node.
– *Improper structuring*: an AND-join receives input that early started from an XOR-split.

These types of patterns are shown in figure 3, where sequence flow edges labeled with 1 show a trivial representation of the multiple source pattern. Edges labeled with 2 show both patterns of loop and improper structuring, indicating that these patterns are not disjoint.

3.3 Limitations and Assumptions

To highlight the key concepts, we limit our work on process models containing only AND- and XOR- gateways. We leave the—more complex— discussion related to OR-joins for future work. We also assume that the queried process models do not contain implicit splits or merges. This means, that the pattern shown in figure 4 (a) must not appear; instead the explicit split shown in figure 4 (b) must be used. The similarity is stated in the BPMN specification document [1, p.111].

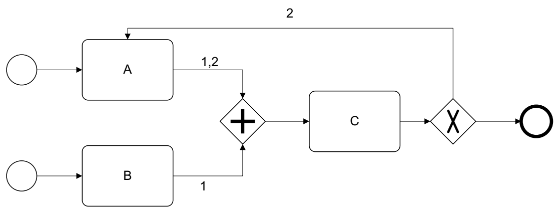


Fig. 3. Faulty Process Model indicating the Types of Deadlock Patterns

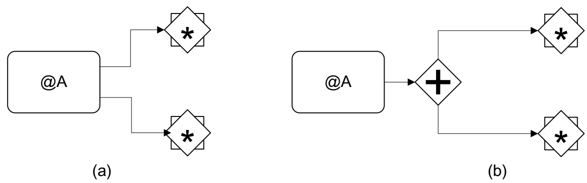


Fig. 4. Implicit versus explicit AND-split

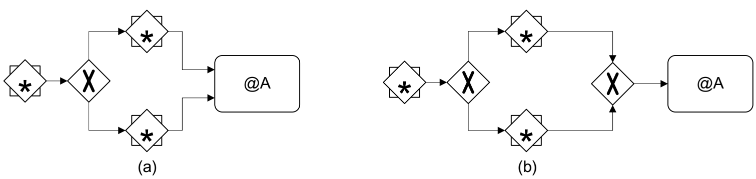


Fig. 5. Implicit versus explicit XOR-join

Unfortunately, implicit merges may not be mapped directly to explicit representations. The equivalent explicit representation depends on the type of the preceding split. Figure 5 shows the specific case, where an implicit merge is interpreted as an XOR-split (for more details about other scenarios please refer to the specification document). Another issue regarding multiple start events is given when all outgoing flows from these start events are leading to the same activity. In this case, there is an implicit AND-join node in front of the activity and all start events are necessary to start execution. To remove this ambiguity, we enforce an explicit AND-join.

4 Deadlock Queries

In this section we present a set of queries that detect the different deadlock patterns. Some of these queries are direct representation of the patterns, whereas others need more details to ensure correct capturing of the patterns.

4.1 Loop Deadlock Pattern

Figure 6 shows the corresponding query in BPMN-Q that declaratively describes the loop pattern. We have numbered the nodes to ease the explanation. Nodes 1

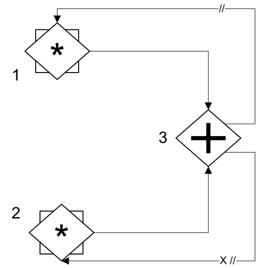


Fig. 6. Loop Deadlock Query

and 2 are called generic nodes that at query evaluation can match to any type of nodes in the process model. Both nodes 1, 2 represent input points to the AND-join (node 3). The path edge starting from node 3 back to node 1 represents an execution path from the AND-join back to node 1 (loop). To be sure that the resolved paths only reach a subset of the input points to node 3, we have put an extra constraint that is represented in the negative path from node 3 to node 2. This means that for some input point(s) of node 3 we fail to find that loop; i.e. the loop covers only a subset of the input points of the AND-join and hence a deadlock occurs.

4.2 Multiple Source Deadlock Pattern

With this pattern we address the case when a process has multiple start points (start event) which are later on synchronized. The query in figure 7 indicates two sources (nodes 1, 2), which are independent of each other. Each of them provides input to a subset of input points (nodes 3,4) to the AND-join (node 5). If we assume that only one of the start events is required to instantiate the process—such as in BPMN—a deadlock occurs at the AND-join.

4.3 Improper Structuring

The mapping of the first two deadlock pattern to BPMN-Q was almost straightforward. When we consider the third case of deadlocks, the mapping is not straightforward. If we consider a direct mapping from the pattern description to

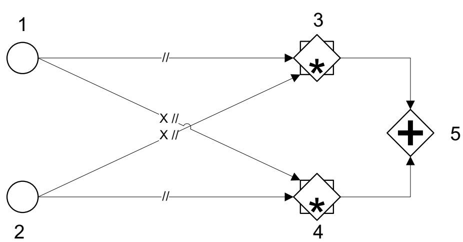


Fig. 7. Multiple Source Query

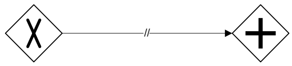


Fig. 8. Direct Description of the Improper Structuring Pattern as a Query

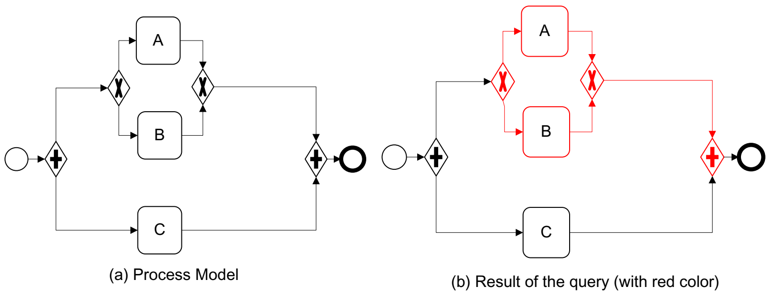


Fig. 9. Example: Deadlock-free Process Model and Query Result

a query, we would come out with a query like the one shown in figure 8. If we apply this query to the process model shown in figure 9(a), however, it would result in figure 9(b), showing a match. Although the query found a match in the process model, the model actually is not suffering from deadlocks. The mistake lies in the fact that there is *no* improper structuring here because the AND-join gate is not the match of the XOR-split, since there is an XOR-join in-between.

To solve the issue, we need to modify this query to address patterns where an AND-join node is considered as the match to an XOR-split and whatever lies in-between is properly structured. We look for patterns that start with an XOR-split and end with an AND-join and what lies in-between is properly structured. The different combinations of proper structures in-between an XOR-split and an AND-join are given by:

- Balanced XOR-splits/joins and AND-splits/joins.
- Balanced XOR-splits/joins.
- Balanced AND-splits/joins.
- Sequence of nodes that does not contain gateways.

In the following subsections we will discuss each case along with the query that expresses it.

Balanced XOR-splits/joins and AND-splits/joins. The query shown in figure 10 looks for a pattern where there is some XOR-split (node 1) and an AND-join (node 6) and in-between there are balanced structures of XOR-splits/joins and AND-splits/joins. To illustrate that the XOR-split (node 1) is not matched by an XOR-join (node 3), we stated that there is some node in the process graph (node 2) that follows the XOR-split (sequence flow edge between nodes 1, 2) but this node (node 2) is not enclosed in an execution path to the XOR-join (negative path edge between nodes 2, 3). Also we account for the existence of

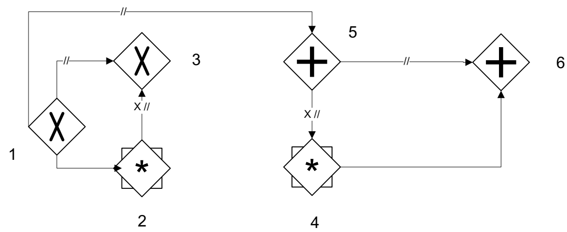


Fig. 10. Improper Structure with in-between Balanced XOR/AND-splits/joins

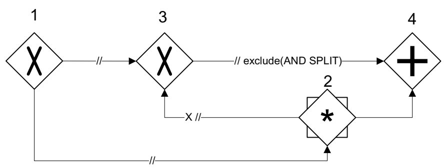


Fig. 11. Improper Structure with in-between Balanced XOR-splits/joins

balanced AND-split/AND-join. In the same way we need to prove that there is *no* AND-split (node 5) that is considered as a match to the AND-join (node 6), by showing that there is some input node (node 4) to the AND-join (sequence flow edge between nodes 4, 6) that is not included in an execution path starting at the AND-split (negative path edge between nodes 5, 4).

Balanced XOR-splits/joins. In this case, we focus on only detecting balanced XOR-gateways that lie between a source XOR-split node and a target AND-join node. Figure 11 is the query that detects this type of deadlock. The detection of the unmatched XOR-split is similar to figure 10. The difference is, that an execution path from the XOR-join (node 3) to the AND-join (node 4) excludes any AND-split nodes (path edge between nodes 3, 4).

Balanced AND-splits/joins. Here, only balanced AND-splits/joins are in-between a source XOR-split and a destination AND-join. Figure 12 shows the corresponding query. We follow the same approach as with balanced XOR-split/joins, but this time we need to show that we fail to find and AND-split (node 2) that acts as a match to the AND-join (node 4). We express this no match by finding a node (node 3) which is an input to the AND-join that is not enclosed by the AND-split (negative path edge between nodes 2, 3). We stress that the path from the XOR-split (node 1) to the AND-split (node 2) contains no XOR-joins (path edge between nodes 1, 2).

Sequence of nodes. The fourth case occurs when an XOR-split is matched by an AND-join with only a sequence of activities or intermediate events in-between (without any gateways). To express this as a query, we use the *exclude* property of path operator in BPMN-Q. The visualization of the query is the same as the one in figure 8, but the path exclude condition is *Type(Gateway)*.

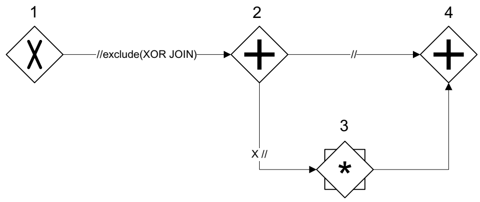


Fig. 12. Improper Structure with in-between Balanced AND-splits/joins

Finally, we need to state that the different queries have to be applied in a certain order. First, the query shown in figure 10 needs to be applied. If it does not find a match, either the query shown in figure 11 or 12 are applied. If both do not find a match, finally the query from figure 8 with the additional exclude condition of gateways should be applied. If all these queries fail to find a match, the process model is free of the kinds of deadlocks we discussed.

5 Example

Figure 13 shows a sample process model that should be examined. When we check queries in the order specified in the end of section 4, the query in figure 10 is the one that finds a match in the process as follows: the XOR-Split (node 1) in the query graph is bound to XOR-Split 1 in the process graph. Node 1 did not bind to XOR-Split 2 because BPMN-Q failed to find a node (in process graph) that is a successor of XOR-Split 2 and in mean time does not have a path to

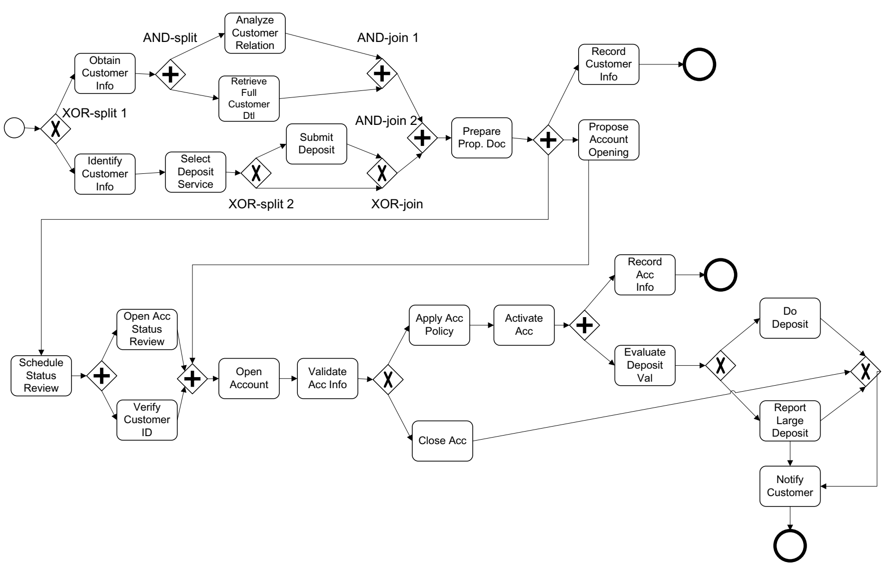


Fig. 13. Process Model suffering from a deadlock

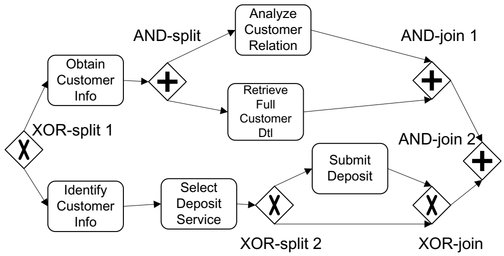


Fig. 14. Query of figure 10 matched to process model in figure 13

XOR-Join. Generic node (node 2) is bound to activity *Obtain Customer Info* as it is a direct successor to the XOR-Split (by the sequence flow edge) and it has no path to an XOR-join node (negative path edge between node 2 and 3 in the query graph). In turn, node 3 in the query is bound to the XOR-Join in the process graph. Generic Node 4 is bound to the XOR-join as this node satisfies the constraint of being a predecessor of an AND-Join node (AND-Join 2) and there is no execution path from an AND-Split to it. Node 5 is bound to the AND-Split node and the AND-Join node (node 6) is bound to AND-Join 2 as we pointed shortly before.

To visualize the query, we reproduced the process model from figure 13 in figure 14. It contains all nodes in the match of the query where the source of the mismatch is XOR-Split 1 and the destination is the AND-Join 2.

6 Conclusion

The paper discussed a novel approach to detect deadlocks in process models via querying them with certain deadlock patterns. In particular, we discussed the translation from deadlock patterns into BPMN-Q queries. This translation required a deeper understanding of the BPMN as well as the BPMN-Q semantics. The results, however, may be applied by process designers without any deeper understanding of BPMN-Q. If a BPMN-Q deadlock query produces a non-empty subset of a process model's graph, this sub-graph directly highlights the problematic areas of the process model. Hence, the process designer is able to directly fix the problems that occur. Regard practical feasibility, our approach has the advantage of requiring low computational effort. The drawback, however, is the incompleteness of the results. If a query matches, a deadlock is found. If no deadlock query matches a given process model at all, this doesn't guarantee that the model is deadlock free. What has been proven instead, is that certain types of deadlocks do not occur in the process model.

Future work will focus on expanding the deadlock detection patterns. We will add support for additional types of unwanted behavior in process models, like for instance lifelocks. We also plan to investigate the required efforts for a larger set of process models, where we additionally plan to compare the results with traditional soundness investigations [2].

References

1. Business Process Modeling Notation (BPMN) Specification, Final Adopted Specification. Technical report, OMG (2006)
2. Aalst, W.: Verification of Workflow Nets. In: Azéma, P., Balbo, G. (eds.) ICATPN 1997. LNCS, vol. 1248, pp. 407–426. Springer, Heidelberg (1997)
3. Aalst, W., ter Hofstede, A., Weske, M.: Business Process Management: A Survey. In: van der Aalst, W.M.P., ter Hofstede, A.H.M., Weske, M. (eds.) BPM 2003. LNCS, vol. 2678, pp. 1–12. Springer, Heidelberg (2003)
4. Awad, A.: BPMN-Q a Language to Query Business Processes. In: Proceedings of the 2nd International Workshop on Enterprise Modelling and Information Systems Architectures, EMISA 2007 (October 2007)
5. Keller, G., Nüttgens, M., Scheer, A.: Semantische Prozessmodellierung auf der Grundlage "Ereignisgesteuerter Prozessketten (EPK)". Technical Report 89, Institut für Wirtschaftsinformatik, Saarbrücken (1992)
6. Martens, A.: Analyzing Web Service based Business Processes. In: Cerioli, M. (ed.) FASE 2005. LNCS, vol. 3442, pp. 19–33. Springer, Heidelberg (2005)
7. Onoda, S., Ikkai, Y., Kobayashi, T., Komoda, N.: Definition of deadlock patterns for business processes workflow models. In: HICSS 1999: Proceedings of the Thirty-second Annual Hawaii International Conference on System Sciences, Washington, DC, USA, vol. 5, pp. 50–65. IEEE Computer Society, Los Alamitos (1999)
8. Puhlmann, F., Weske, M.: Investigations on Soundness Regarding Lazy Activities. In: Dustdar, S., Fiadeiro, J.L., Sheth, A.P. (eds.) BPM 2006. LNCS, vol. 4102, pp. 145–160. Springer, Heidelberg (2006)
9. Sadiq, W., Orlowska, M.E.: Applying graph reduction techniques for identifying structural conflicts in process models. In: Jarke, M., Oberweis, A. (eds.) CAiSE 1999. LNCS, vol. 1626, pp. 195–209. Springer, Heidelberg (1999)
10. van Dongen, B.F., Mendling, J., van der Aalst, W.M.P.: Structural patterns for soundness of business process models. In: EDOC 2006: Proceedings of the 10th IEEE International Enterprise Distributed Object Computing Conference (EDOC 2006), Washington, DC, USA, pp. 116–128. IEEE Computer Society, Los Alamitos (2006)

Employee Competencies for Business Process Management

Stefan Eicker, Jessica Kochbeck, and Peter M. Schuler

University of Duisburg-Essen
Faculty of Economics: Chair of Business Informatics and Software Engineering
Universitätsstraße 9, 45141 Essen, Germany
{Stefan.Eicker,Jessica.Kochbeck,Peter.Schuler}@icb.uni-due.de

Abstract. Business process management (BPM) is an approach which empowers companies to react flexibly to new market situations. The main goal of BPM is to improve efficiency and effectiveness of value-adding business processes. The changes caused by globalization do not only concern organization, technologies and processes, but also people. Employee competencies can be crucial competitive advantages.

The need for new specialized and competent personnel in BPM becomes apparent from the definition of new roles, such as "Chief Process Officer" (CPO). Field reports and surveys reveal that role concepts of BPM have so far not been completely established, due to a lack of appropriate structures or due to resistance within the companies.

This article considers and analyzes the success factor *employee competencies* in matters of the implementation of BPM in companies. For this purpose, competencies which are necessary for the roles in BPM are identified. Moreover, a classification method for the definition of role profiles is developed.

Keywords: Business Process Management, Measurement, Inter-Organizational Processes, Human Factors.

1 Introduction

Nowadays companies are facing an increasing stress of competition. They have to cope with shorter product lifecycles, rising customer demands, quicker technological developments and higher cost pressure. Classical production factors such as labor and capital are losing some of their significance. In order to create strategic competitive advantages, companies have to concentrate on their core competencies, which are significantly influenced by the skills and the knowledge of their employees.

To remain competitive despite the consequences of globalization, companies have to consider change as a continuous challenge. Besides organization, technologies and processes, the changes also affect the employees of the company.

One chance for companies to react flexibly to new situations is the targeted management of business processes. The main goal of business process management is to increase efficiency and effectiveness of companies by improving business processes

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 251–262, 2008.

and thus to increase the company value. The implementation of business process management is accompanied by a changeover to process-orientation, leading to modifications in the organizational and operational structure of a company. New roles and responsibilities have to be defined. The operational structure is affected through changes in management and operation as new roles and responsibilities implicate new managerial authorities [1]. For the employees, change implies continuous learning in order to tackle new challenges and tasks. The focus of this article is on the success factor *employee competencies* because these empower companies to achieve competitive advantages.

Process-orientation is meanwhile widespread and of high importance. Nevertheless numerous companies did not name a designated executive officer for the management of business processes [2].

This article targets to answer the following questions which are derived from the above-mentioned aspects:

- Which competencies are necessary to meet the demands of the new roles in business process management?
- How can suited employees be identified?

In chapter 2 the different roles in business process management are described. In the following, the developed classification method for the definition of the levels of competencies will be introduced (chapter 3). Afterwards, the particular competencies are defined (chapter 4). In chapter 5 the competence framework developed for business process management including the competencies derived will be outlined. By means of this framework, the job specifications and occupational aptitude of the employees can be recorded and compared.

2 Roles in Business Process Management

In the following the classification of roles according to SCHMELZER AND SESSELMANN will be introduced. These authors define six roles in business process management[1]. At the implementation stage of BPM, the task of the business process management *project leader* (BPM project leader) is the most responsible one. He may be supported by a *process consultant*. While the business process is executed, the *process owners* have the most important task because they are responsible for a whole process or for a subprocess. Other roles are the *process coordinator*, the *process controller* and the *process staff* [8]. Their responsibilities are shown in Table 1. The overall responsibility is taken by the top management. However, the management has no direct control over the *process staff* because they are reporting to the *process owner* [6].

The *project leader* and the *process consultant* are deployed mainly at the implementation phase of business process management. The people in the roles of the *process coordinator*, the *process owner*, the *process controller* and the *process staff* are responsible for the operation of the business process management [8].

[1] Further examples for role definitions can be found in [3], [4], [5], [6], [7].

Table 1. Roles in business process management[2]

BPM role	Responsibilities
Project leader	Responsible for the implementation of business process management
Process consultant	Conceptual and methodical assistance during the implementation phase
Process coordinator	Responsible for the advancement and the integration of the whole BPM system
Process owner	Responsible for target achievement and improvement of business (sub)processes
Process controller	Responsible for operation and enhancement of process controlling
Process staff	Responsible for the operation of process steps and activities

Fig. 1 shows the roles in business process management and their connections. *Project leader* and *process consultant* are coloured brighter because their functions are only important in some phases of the business process management circle. The dashed arrows represent the one-time appointment of the roles at the implementation phase of business process management. The solid arrows indicate the connection of the roles that are involved in the operation of business process management.

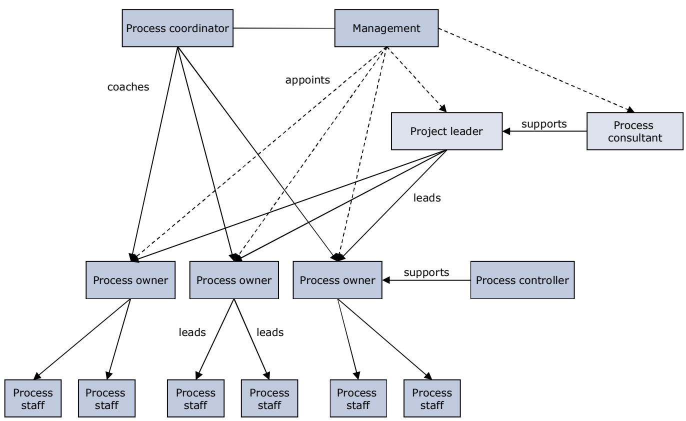


Fig. 1. Relations between BPM roles

In order to fill the various roles in business process management in the best possible way, in a first step a requirements analysis has to be made, i.e. it has to be defined which competencies are necessary for the fulfilment of the tasks. There are different kinds and levels of competencies required for business process management.

[2] According to [8].

These have been identified based on different references[3] and job advertisements and will be focused on in the following. The employee competencies – covered in detail in chapter 4 – are: experience, expertise, determination, conceptual and analytic intellectual power, organizational skill, assertiveness, capacity for teamwork, communication skill, conflict management ability, customer orientation as well as leadership.

3 Classification Method for Levels of Competencies

Qualifications can be measured and proved through references, certificates and credentials. In case of skills which are acquired in informal learning processes, this kind of proof is generally not possible. A distinction in rating scales with parameter values e.g. ranging from "weak" to "strong" are too broad and arbitrary for an exact indication. Thus, they are insufficient for a precise classification. Therefore, competencies are measured by means of a multi-level competence scale [13]. In order to assess subjective characteristics objectively, a precise scale identifying various levels of competence and maturity is essential. The classification shows that an employee on a high level in the company hierarchy must comply with different requirements and must have different competencies than a member of a lower level in the company hierarchy [14].

Competencies are graded in levels, either on the basis of a numerical scale[4] or of a verbal schema[5]. In order to develop a classification scale, both approaches will be combined: In addition to a numerical value marking each level, a verbal description is given. The classification scale is based on an ordinal scale, i.e. the levels of competence are ranked. Negative competencies are not included in this scale, because existing competencies are always positive [20]. If a certain level is reached, this implies that the person also meets the criteria of the lower levels. If one competence is not required in the job specification, the respective field is left blank.

The following level definitions are still very abstract. In chapter 4, these level definitions will be applied to the particular competencies, referring to observable behaviour and skills.

Level 1: Sufficient
The competencies meet the standard. They come up to the basic minimum requirements.

Level 2: Satisfactory
The competencies come up to the basic requirements. The standard is completely met.

Level 3: Good
The competencies exceed the standard. They are well distinctive and meet with the basic requirements over a longer period.

---

[3] see [3], [4], [5], [6], [7], [8], [9], [10], [11], [12].
[4] see [15], [16], [17].
[5] see [18], [13], [19].

Level 4: Very good
The competencies are above average. The requirements specification is completely met.

Level 5: Excellent
The employee has special distinctive competencies exceeding the job specification.

The next figure illustrates the five abstract level definitions. The same layout will be used for the particular competence levels in chapter 4.

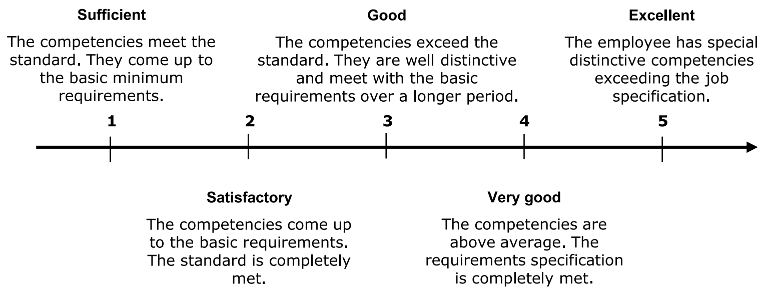


Fig. 2. Definition of abstract competence levels

4 Catalog of Competencies

In this section, the derived competencies required in business process management are listed and described. According to SPENCER AND SPENCER, competencies are described on the basis of three to six attitudes which can often be observed at work day-to-day [21]. In literature, there are different descriptions of competence levels. Below, five competence levels are assigned to each kind of competence required in business process management[6], corresponding to the classification scale developed in chapter 3. Each level describes a working manner referring to the person's attitude. STEINMANN AND SCHREYÖGG describe such a classification scale as a behaviour expectation [24][22]. When measuring employee competencies, the level reached by the person is assessed and afterwards recorded in the corresponding competence framework. The resulting occupational aptitude can be compared to the job specification. The bigger the overlapping, the better is the person suited for the position to be filled. Besides information on possible competence deficits, information on so far unknown and unused competencies can be obtained. Thus it might be advisable to entrust an employee with a different task, in accordance with his qualifications and skills, ensuring that the internal resources in the company are used in the best possible way.

Experience
Competence increases through experience gathered at work. The more professional experience is gathered, the better knowledge with respect to technical, methodical and

[6] According to [14], [19], [21], [22], [23].

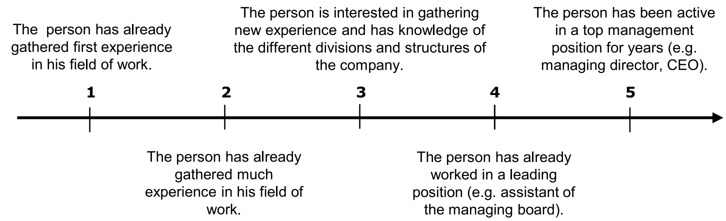


Fig. 3. Levels of competence: Experience

social areas [13] can be achieved. Fig. 3 shows the five particular competence levels for the competence *experience*.

Expertise

Special knowledge such as language skills or programming knowledge is called expertise. It can be acquired through training measures [13]. Fig. 4 shows the five particular competence levels of *expertise*.

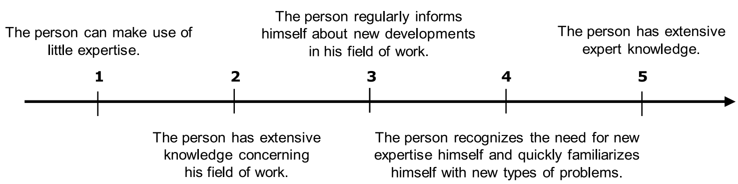


Fig. 4. Levels of competence: Expertise

Determination

Managers are responsible for the implementation of the strategic business objective defined by the top-management. It is the task of managers to make the employees aware of the objectives [25]. Fig. 5 shows the levels of the competence *determination*.

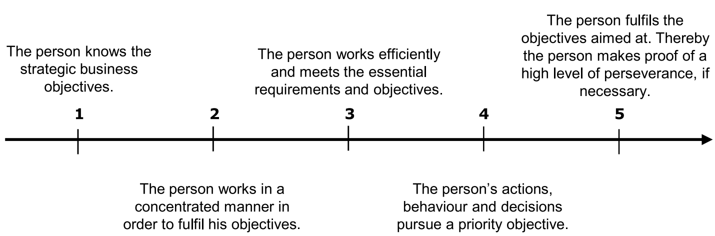


Fig. 5. Levels of competence: Determination

Conceptual and analytical intellectual power

Different tasks require the person, e.g. the process owner, to think and work in a structured and methodical manner. The person works very carefully and, even in

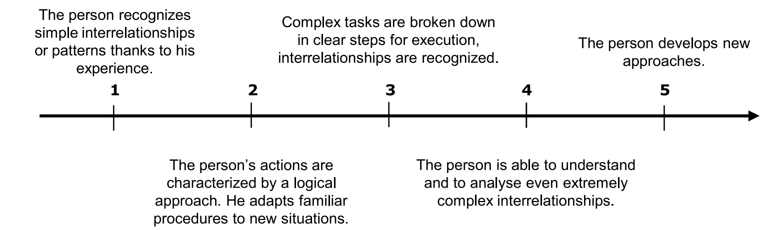


Fig. 6. Levels of competence: Conceptual and analytical intellectual power

complex situations, he does not lose track and acts analytically. The particular competence levels of *conceptual and analytical intellectual power* can be found in Fig. 6.

Organizational skill

In the case of unregulated workflows, it is essential that the employees have a high level of organizational skill. Thus they must be able to recognize coherences and to structure their work [25]. The following Fig. 7 shows the five competence levels of *organizational skill*.

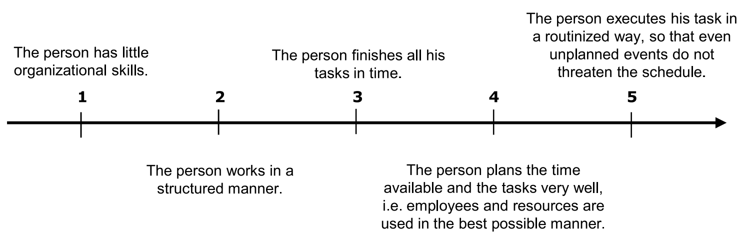


Fig. 7. Levels of competence: Organizational skill

Assertiveness

The person is able to convince his colleagues and employees of his ideas and objectives. He is willing to implement them, even in the case of resistance [13]. Assertiveness is consequent orientation of one's actions to the target desired [26]. The five competence levels of *assertiveness* are illustrated in Fig. 8.

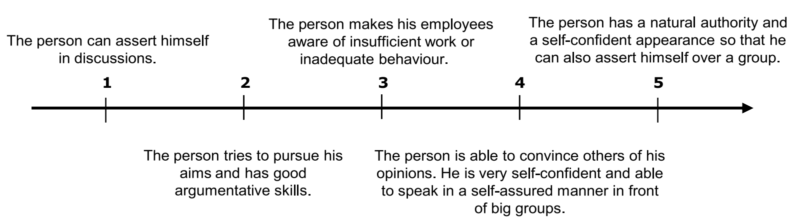


Fig. 8. Levels of competence: Assertiveness

Capacity for teamwork

Persons who have capacity for teamwork use their experience and knowledge to play a part in a team in order to reach common targets. Among their skills, there is willingness to compromise and openness to suggestions and arguments from team members. In order to work successfully in a team, the persons must have equal rights and be tolerant [13]. Fig. 9 details the particular levels of the competence *capacity for teamwork*.

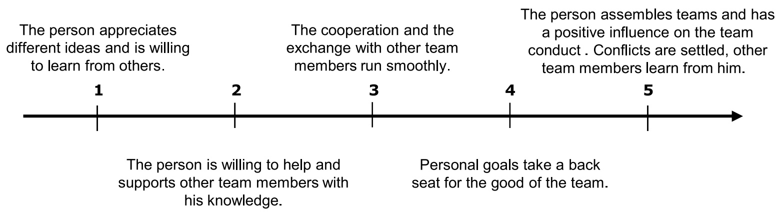


Fig. 9. Levels of competence: Capacity for teamwork

Communication skill

The person has the talent to communicate clearly, to listen attentively and to distinguish important from unimportant information during a conversation. An important characteristic is the correct interpretation of gesture and other nonverbal signals which help to understand the conversational partner and to be responsive to his concerns [15]. The competence levels of *communication skill* can be found in Fig. 10.

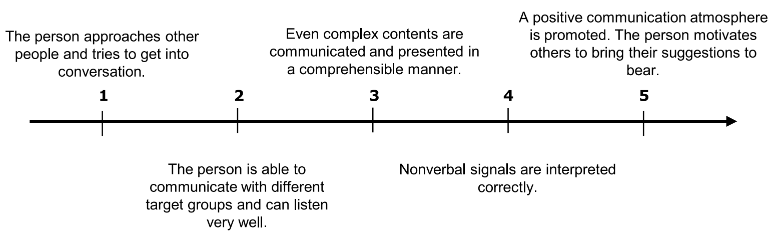


Fig. 10. Levels of competence: Communication skill

Conflict management ability

Whenever several people work on a common task as a team, inevitably conflict situations arise in which different opinions and ideas are expressed. In order to avoid escalation and negative influence on the project, it is important to behave correctly in a conflict situation. An employee who is able to deal well with conflicts is likely to take and give criticism in an objective and constructive manner. Criticism ought to be viewed as a chance to reconsider one's own opinion. Fig. 11 shows the five competence levels of the *conflict management ability*.

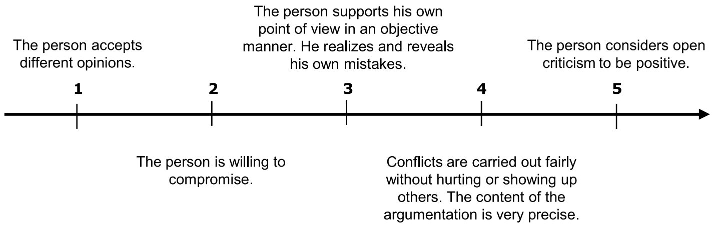


Fig. 11. Levels of competence: Conflict management ability

Customer orientation

Customer orientation means that the customer is the centre of corporate decisions and actions. The aim consists in reaching high customer satisfaction in order to establish durable customer relations. Fig. 12 shows the competence levels of *customer orientation*.

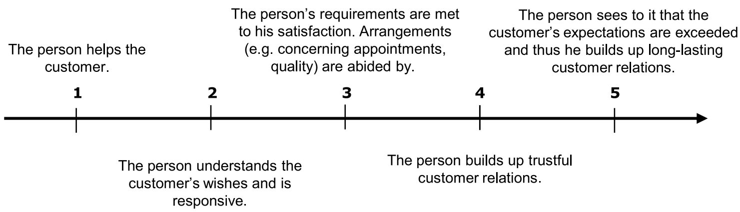


Fig. 12. Levels of competence: Customer orientation

Leadership

Superiors or project managers have to be informed about the knowledge and skills of their employees so that they can use them in a targeted way. Professional training is to be offered and feedback referring to the work performed is to be given. Independent and self-responsible actions are to be promoted. Decisions must be objective and comprehensible to the employees [13]. The five particular competence levels of *leadership* are shown in Fig. 13.

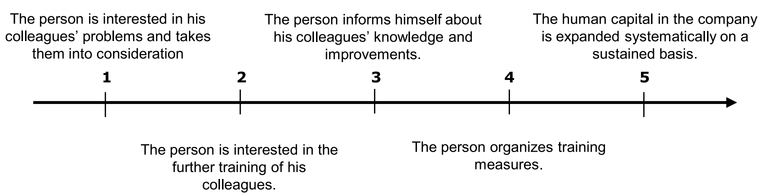


Fig. 13. Levels of competence: Leadership

5 Competence Framework for Business Process Management

The results from assessment methods for employee competencies, such as assessment centres or interviews, can be recorded in an aptitude profile. From these profiles, it is possible to derive in which area competencies are at a high or at a low level so that corresponding measures can be taken [22]. Through comparison with the job specification, the occupational aptitude of different candidates can be ascertained and compared.

The competencies required in business process management are summarized in the competence framework described in the following (see Table 2). The competencies are grouped into specialist, character, methodical, social and entrepreneurial competencies. Each line shows the five point scale corresponding to the competence levels defined beforehand.

Table 2. Exemplary competence framework

Role	Process owner				
Responsibility	Responsible for designing, operating and improving the business process				
	1	2	3	4	5
Specialist competencies					
Experience					
Expertise					
Character competencies					
Determination				X	
Methodical competencies					
Intellectual power				X	
Organizational skill			X		
Social competencies					
Assertiveness				X	
Capacity for teamwork			X		
Communication skill			X		
Conflict management ability				X	
Entrepreneurial competencies					
Customer orientation					X
Leadership				X	

= Job specification, X = Aptitude profile

Competencies in the table are structured as follows: At first, the role of the employee is listed, followed by a description of his core tasks. From here, the competencies required are derived and recorded in the job specification. In a first step, specialist competencies are listed. These competencies (such as foreign language skills, modelling experience with particular tools, stays abroad, legal knowledge etc.) correspond to a certain position instead of a role and can be defined exactly. In a next step, the character and methodical competencies required for the role are listed. Social competencies and finally entrepreneurial competencies complete the framework.

The competence framework illustrated, is exemplified by the role of the *process owner*. The specialist competencies (marked by half-filled fields) have to be defined in accordance with the circumstances (e.g. international experience, ARIS skills and Spanish language skills).

The job specification of the role *process owner* is coloured in grey. Crosses mark the exemplary aptitude profile of an employee.

6 Conclusion

For the filling of new roles in business process management, suited employees must be identified; the developed method for the definition of competence profiles supports this procedure. Experience reports and surveys show that the role concepts of business process management presented have so far not been implemented completely due to a lack of structures or resistance in the company. Nevertheless, already new roles such as CPO are defined. This shows the need for competent personnel.

Job specifications are subject to continuous change because of globalized markets and process and product innovations. Thus they have to be adapted regularly. The competence framework presented in this article can be supplemented with new competencies or reduced depending on demands on the employees. Moreover it is possible to fine-tune the model through different weighting of the particular competencies. The more differentiated and comprehensive the roles are, the more difficult it is to analyze and record the qualifications required, so that the job specification does not consist of an unmanageable number of competencies.

In the context of change management, the method developed can be used as a starting point in order to not only identify competence deficits of employees, but also to prepare them for their future tasks.

References

[1] Mischak, R.F.: Business Reengineering – Der Weg vom funktions- zum prozessorientierten Denken im Unternehmen. In: Berndt, R. (ed.) Business Reengineering, pp. 3–17. Springer, Berlin (1997)

[2] Institut für Softwaretechnik und Interaktive Systeme: Studienergebnisse Status Quo Geschäftsprozessmanagement 2005 – Deutschland (2005), http://www.ifs.tuwien.ac.at/gpm-studie/2005/GPM-Studie-2005_Ergebnisse-Deutschland.pdf

[3] Osterloh, M., Frost, J.: Prozessmanagement als Kernkompetenz. Wie sie Business Reengineering strategisch nutzen können. Gabler, Wiesbaden (2003)

[4] Fischermanns, G.: Praxishandbuch Prozessmanagement. Ibo Schriftenreihe, Band 9. Verlag Dr. Götz Schmidt, Gießen (2006)

[5] Best, E., Weth, M.: Geschäftsprozesse optimieren. Gabler, Wiesbaden (2005)

[6] Hammer, M., Champy, J.: Reengineering the Corporation: A Manifesto for Business Revolution. Nicholas Brealey Publishing, New York (2001)

[7] Neumann, S., Probst, C., Wernsmann, C.: Continuos Process Management. In: Becker, J., Kugeler, M., Rosemann, M. (eds.) Process Management A Guide for the Design of Business Processes, pp. 233–250. Springer, Berlin (2005)

[8] Schmelzer, H.J., Sesselmann, W.: Geschäftsprozessmanagement in der Praxis. Kunden zufrieden stellen – Produktivität steigern – Wert erhöhen. Hanser Verlag, München (2006)

[9] Brodbeck, P.: Einführung von Prozessmanagement. In: Hirzel, M., Kühn, F. (eds.) Prozessmanagement in der Praxis. Wertschöpfungsketten planen, optimieren und erfolgreich steuern, Gabler, Wiesbaden, pp. 171–181 (2005)

[10] Becker, J., Berning, W., Kahn, D.: Project Management. In: Becker, J., Kugeler, M., Rosemann, M. (eds.) Process Management A Guide for the Design of Business Processes, pp. 13–40. Springer, Berlin (2003)
[11] IDS Scheer AG: Business Process Report (2006), http://www.ids-scheer.com/international/german/press/?template=formular_email_bpr
[12] Gaitanides, M., Scholz, R., Vrohlings, A., Raster, M.: Prozessmanagement. Carl Hanser, München (1994)
[13] North, K., Reinhardt, K.: Kompetenzmanagement in der Praxis. Mitarbeiterkompetenzen systematisch identifizieren, nutzen und entwickeln. Gabler, Wiesbaden (2005)
[14] Gerhardt, T., Ritter, J.: Management Appraisal. Kompetenzen von Führungskräften bewerten und Potenziale erkennen. Campus, Frankfurt am Main (2004)
[15] Bühner, R.: Mitarbeiterkompetenzen als Qualitätsfaktor. Strategieorientierte Personalentwicklung mit dem House of Competence. Hanser, München (2004)
[16] Paschen, M., Weidemann, A., Turck, D., Stöwe, C.: Assessment Center professionell. Hogrefe, Göttingen (2005)
[17] Olesch, G., Paulus, G.J.: Innovative Personalentwicklung in der Praxis. Mitarbeiter-Kompetenz in der Praxis. Mitarbeiter-Kompetenz prozessorientiert aufbauen. C.H. Beck, München (2000)
[18] Mathis, R., Jackson, J.H.: Human Resource Management. South-Western, Ohio (2003)
[19] Sommerhoff, B.: Mitarbeiterbeurteilung. Leistung messen, Mitarbeiter fördern, Personal entwickeln. Verlag Moderne Industrie, Landsberg (1999)
[20] Erpenbeck, J.: KODE – Kompetenz-Diagnostik und -Entwicklung. In: Erpenbeck, J., von Rosenstiel, L.: Handbuch Kompetenzmessung. Erkennen, verstehen und bewerten von Kompetenzen in der betrieblichen, pädagogischen und psychologischen Praxis, Schäffer-Poeschel, Stuttgart, pp. 365-375 (2003)
[21] Spencer, L.M., Spencer, S.M.: Competence At Work. John Wiley & Sons Inc., Canada (1993)
[22] Stolzenberg, K., Heberle, K.: Change Management. Veränderungsprozesse erfolgreich gestalten – Mitarbeiter mobilisieren. Springer, Heidelberg (2006)
[23] Jost, W.: Vom CIO zum CPO. In: Harvard Business Manager, September 2004, pp. 88–89 (2004)
[24] Steinmann, H., Schreyögg, G.: Management. Grundlagen der Unternehmensführung-Konzepte – Funktionen – Fallstudien. Gabler, Wiesbaden (2000)
[25] Heyse, V., Erpenbeck, J.: Kompetenztraining. 64 Informations- und Trainingsprogramme. Schäffer-Poeschel, Stuttgart (2004)
[26] Niermayer, R.: Soft Skills – Das Kienbaum Trainingsprogramm. Haufe, München (2006)

Enterprise Web Services and Elements of Human Interactions

Peter Géczy, Noriaki Izumi, Shotaro Akaho, and Kôiti Hasida

National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba and Tokyo, Japan

Abstract. We present guidelines for improving interactivity between web services and human users. The guidelines advocate matching the essential characteristics of human behavior in electronic environments to the human-service interactions. The study examines interactions of knowledge workers on a large corporate intranet portal and infers relevant implications to web services. The results indicate that effective human interactions with web services should not require more than five user actions, on average. Active interactions should have dynamics in the range of seconds, and passive interactions should not demand more than seven minutes of human attention. Complex business processes should be segmented and implemented via logical compacts of interactive sub-processes. The number of sub-processes should not exceed three.

1 Introduction

Service-oriented architecture and computing principles [1],[2] are significantly transforming the business information technology landscape [3]. Separating service functionality from its implementation enables composition of services across platforms and organizations [4]-[6]. Resulting data exchange and aggregation over organizational boundaries raises important privacy concerns [7],[8]. It also presents challenges for interoperability of services.

Service-to-service interactions and their modeling have been widely addressed by researchers and practitioners [9],[10]. However, investigation of the human-service interactions is lacking. Elucidation of human dynamics in web environments [11],[12] has been attracting substantial attention from corporate spheres. Corporations are constantly looking for novel opportunities to gather behavioral data about their web customers and visitors [13]. They are aiming at converting more web visitors into customers, and accurately targeting their marketing to relevant groups and demographics. Unfortunately, the internal focus on usability of their own intranet portals and services, benefiting the employees, is far behind.

This work presents a pioneering attempt to address the pertinent issues in improving human-service interactions in the organizational environments. We analyzed extensive data containing knowledge worker interactions on a large corporate intranet portal. Analytic results revealed important findings widely applicable to human-service interactions.

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 263–272, 2008.

2 Human-Service Interactions

We introduce the essential terminology, the target and the scope of the investigation, and the approach angle. To highlight the issues under consideration we provide an illustrative example of a business process implemented on a corporate intranet portal. It depicts the rudimentary elements of human-service interactions (HSI).

Categorization of web services can be done according to their composition and interactivity. Compositional perspective distinguishes:

- **Atomic Services:** stand-alone functional components not utilizing functionality of other services.
- **Compound Services:** incorporating functionality of one or more other atomic or compound services.

Analogously, we can categorize web services with respect to their interactivity either with other services or human users:

- **Interactive Services:** interact with other services or human users.
- **Non-interactive Services:** do not interact with human users and/or other services.

Although the service-to-service interactivity has been extensively elucidated by researchers and practitioners [9], the human-service interactivity is largely unexplored. In enterprise and organizational environments the internal portals often provide rich pool of services and resources available to the members. Internal information infrastructure, organization, and functionality should improve work efficiency and productivity. Unfortunately, it is still rarely the case. The internal enterprise portals, despite vast number of resources and services, are often underutilized [14]. It is largely due to the misalignment between human browsing characteristics and management and implementation of business process via web services. Constructive feedback and actionable knowledge on human-service interactions is vital for designers, managers, as well as academic and scientific community.

Human users can interact with the services in various ways. Two generally distinguishable interactive categories are:

- **Passive HSI:** this underlines significant one-way (or biased) interaction pattern: service $\rightarrow$ human, or service $\leftarrow$ human. Humans or services are largely either on the recipient's or supplier's side.
- **Active HSI:** this delineates significant two-way (or balanced) interactive pattern: service $\leftrightarrow$ human.

An example of a passive human-service interaction would be a simple display of information, e.g. in textual or graphical form, where the human user is simply the consumer of the information. Active human-service interactions can be wide ranging; from simple acknowledgment clicks, to multimedia-style interactivity.

Consider an example of an attendance monitoring business process. Many organizations are implementing it in the form of a web service on the organizational

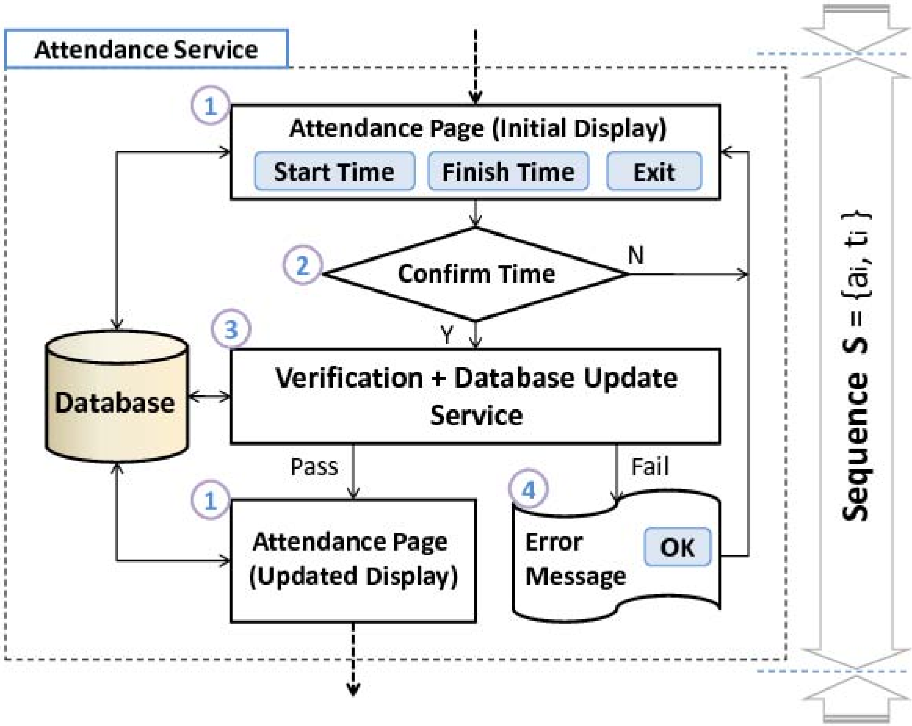


Fig. 1. Simplified attendance service flowchart. The service contains two atomic services (2 and 4) and two compound services (1 and 3). Services 1, 2, and 4 interact with human users, whereas service 3 does not. Human-service interactions are represented by a sequence $S = \{(a_i, t_i)\}$ of actions a_i at the times t_i.

intranet. The users, at the beginning of the work, are required to login into the intranet, locate the attendance service, and record their starting time of work. At the end of the working day similar procedure is required to record their finishing time. Apart from just simply recording the starting and ending times, the service may account for diversity in employee contracts, working hours and style (e.g. fixed or flexible working hours), overtime work, location (e.g. office within the country or overseas locations having different time zones, or even home office), etc. The service should interface with users in a personalized manner fitting their particular situation and working conditions. Even seemingly simple business process implementation via web service may require interoperability with several other services and effective interactions with human users.

The illustrative flowchart of the attendance service is depicted in Figure 1. Users are presented with the initial display of information relevant to their particulars. When clicking the 'Start Time' button, the users are presented with a window containing the recorded current local time, possible time adjustment options, and other functionality. The users can adjust their starting time and applicable options (if required), or simply acknowledge the initially displayed data.

After confirming the information, the service has to verify the data and update database containing working records and possibly preferences. If the verification and database update are successful, the properly updated attendance page is displayed. In case of failure, an error message is shown and users are taken to the adjusted attendance display for re-try or exit. Similar process is repeated when recording the finishing time.

The presented compound attendance service contains four other services:

1. *Attendance Display:* retrieves user's work and preference records, and presents personalized display of attendance information.
2. *Time Confirmation and Adjustments:* gathers time information, offers time adjustment functionality and other optional setting.
3. *Verification and Database Update:* verifies the appropriateness of supplied data and manages secure update of relevant database records.
4. *Error Message Display:* translates the error code into understandable form and offers appropriate suggestions to resolve the problem.

Services 2 and 4 are atomic, whereas services 1 and 3 are compound. Services 1, 2, and 4 interact with human users, but service 3 interacts only with other services. All human interactive services are active since they require dynamic user actions.

Human user interactions with the attendance web service are represented by a sequence of actions executed at the specific times: $S = \{(a_i, t_i)\}_i$, where a_i denotes the action i at the time t_i. It would be nice if we could record all the appropriate actions and times. Unfortunately, this is commonly not possible in organizational intranet setups. However, the majority of user actions result in the factual requests to the web server providing the service. The web servers have capacity to record the requests in a log file. Thus the interactions can be analyzed from the web server log records.

Human interactions in web environments are broader than just interactions with web services. If we narrow the observations to the recorded sequences corresponding to human interactions only with web services, all we could say is how human users interacted with the specific set of services implemented on the studied portal. This would be relatively shallow and of limited practical value since other portals may have different set of services implemented differently. We favor broader observations of human interactions in web environments of which services are part. This enables us to observe the elements of human behavioral aspects that apply to web services in general.

Human behavior in web environments can be segmented according to the user activity and inactivity [15]. Larger behavioral segments of user web interactions — sessions — comprise of smaller behavioral segments; so-called subsequences. Sessions underline more complex user tasks that are divided into several sub-tasks represented by subsequences. Human-service interactions occur at the level of subsequences. In other words, behavioral subsequences are supersets of human-service interaction sequences. This enables us to provide a higher-level perspective on effects and implications of human browsing behavior elements to web services.

3 Intranet Portal Case Study

Presented work uses intranet web log data of The National Institute of Advanced Industrial Science and Technology (Table 1). The institute has a significantly large intranet portal. Its infrastructure consists of six web servers connected to the high-speed backbone in a load balanced configuration. The intranet services and resources are available to the users via channels ranging from optical to wireless connectivity, and support platforms extending to mobile devices.

The institution has a broad network of branches at various locations throughout the country; thus services and resources are decentralized. The portal contains wide spectrum of resources, such as documents in various formats, downloadable software, multimedia, etc. There is also large number of services supporting organizational business processes, cooperation with industry, academia, and other institutes, resource localization; but also networking, blogging, etc. Visible web space exceeds 1 GB, and deep web space is considerably larger, but hard to estimate due to the distributed architecture and altering back-end data.

Table 1. Basic information about data used in the study

Log Records	315 005 952
Resources	3 015 848
Services	855
Sessions	3 454 243
Unique Sessions	2 704 067
Subsequences	7 335 577
Unique Subsequences	3 547 170
Valid Subsequences	3 156 310
Unique Valid Subsequences	1 644 848
Users	$\sim$10 000

Intranet traffic is considerable and generates a rich pool of web log data. The data was, however, contaminated by machine generated traffic (from monitoring software and several other applications) and required extensive preprocessing, and cleaning. The data preparation, processing, filtering, and segmentation to sessions and subsequences are described in [15] and not addressed here. It is worth noting that elimination of machine generated traffic substantially reduced the number of unique valid subsequences, by 53.6% (the primary target of investigation is human interactivity–hence the machine traffic elimination).

4 Exploratory Findings and Web Service Implications

This section presents analysis of service utilization characteristics by knowledge workers as well as their browsing behavior on organizational intranet. Relevant implications pertinent to enterprise-level web services are inferred.

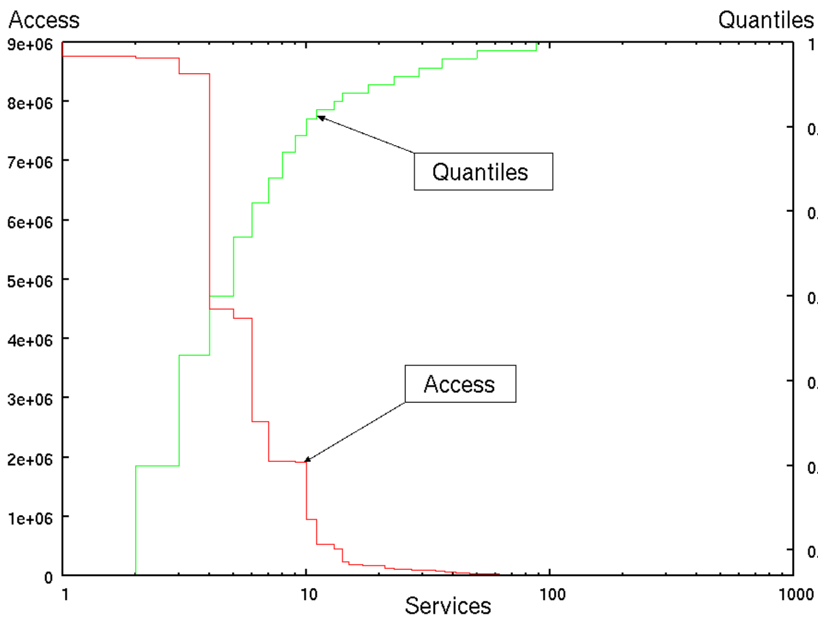


Fig. 2. Service access histogram and quantile characteristics. X-axis is in logarithmic scale. Left y-axis refers to access histogram and right y-axis refers to quantiles.

Relatively few services are frequently used. Histogram and quantile characteristics depicted in Figure 2 indicate that only approximately ten services (out of 855) were frequently used by knowledge workers. Over half of the services were accessed less than ten times. Top three services, i.e. 0.35%, accounted for 50% of use. These were: bulletin board (containing organizational and other announcements), attendance service (recording, verifying, and altering the presence, business trip, and holiday records), and information service from human resources division (finding information about the members of the organization, such as e-mail, phone numbers, location, etc.). This was the actual order. Bulletin boards were accessed more frequently than the attendance service. It indicates that knowledge workers were generally interested in organizational and personal announcements. As shown in Figure 2, the statistics of web services access display evident long tail characteristics.

Frequently accessed services should be easily accessible, highly optimized and personalized. Unfortunately, large number of organizations does not pay sufficient attention to portal analytics, optimization, and personalization of web services. They often use one-fit-all web templates resulting in low portal usability.

Efficient interactive web services should not require more than five user actions. This should be taken into account for both atomic and compound web services. It has been observed from the analysis of knowledge worker web behavior that they divide, on average, their intranet browsing tasks (sessions) into three sub-tasks (subsequences). Each subsequence represents a compact browsing

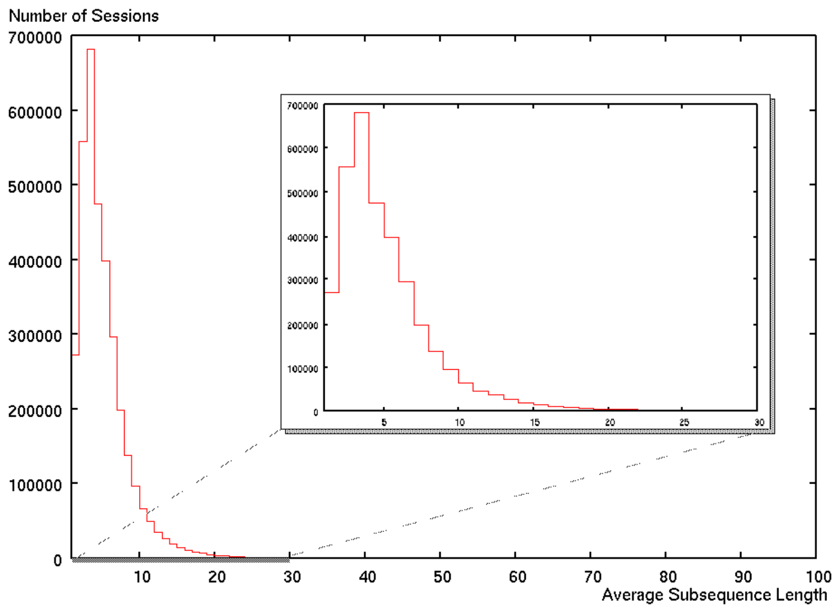


Fig. 3. Histogram of average subsequence lengths in browsing sessions. The detailed view of the initial histogram part is shown in sub-window.

segment. As it can be seen from Figure 3 the peak subsequence length is between two and five. In order to maintain the users' natural browsing flow and segmentation, it is beneficial to limit the number of interactions with web services to approximately five actions. If a business process requires more than five human interactions, it is advisable to implement it as a sequence of logical segments with each requiring less than five human actions.

Implementation of interactive web services should also avoid looping, i.e. when the failure to complete the business process successfully leads to the start of the interactive loop over-and-over again. Interruption management should be implemented into such process [16]. After two or three failures the business process execution should be interrupted and automated assistance services should be invoked. Analogously, the automated assistance services should also adhere to the mentioned principles.

Active human-service interactions should have dynamics within few seconds. The observed average subsequence duration in knowledge worker browsing sessions was 30.68 seconds. If during this duration the users had, on average, five interactions, one interaction had lasted approximately six seconds (on average). Histogram in Figure 4 displays the average subsequence duration statistics in sessions. It is interesting to notice that the peak average subsequence duration was between three and six seconds. Thus the large number of subsequences lasted less than six seconds — which implies approximately one second per interaction. Users exhibited fast-paced dynamics. It should be noted, however, that the active

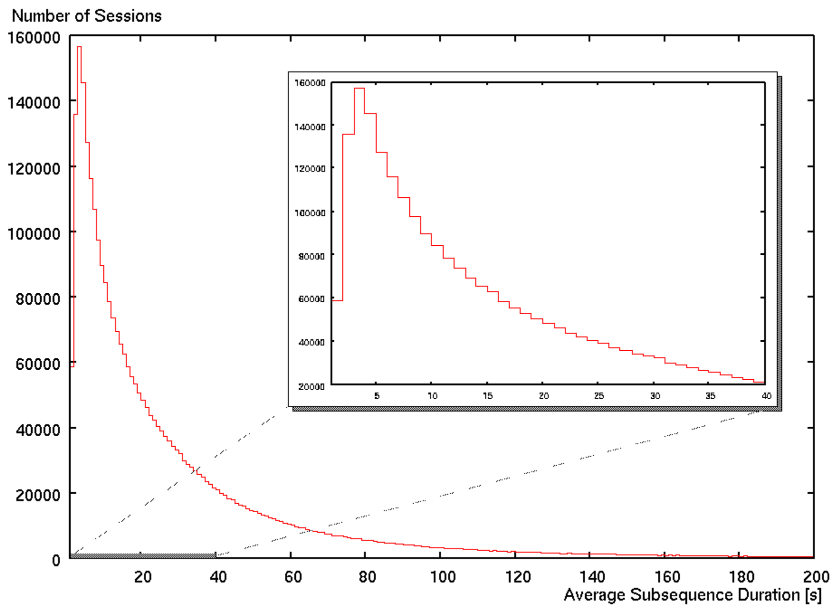


Fig. 4. Histogram of average subsequence durations in browsing sessions. The detailed view of the initial histogram part is displayed in sub-window.

interactions on the studied portal were relatively simple. More sophisticated interactive modes (e.g. multimedia-style) may last longer.

Passive human-service interactions should be limited to approximately seven minutes. Detected average duration of knowledge worker browsing sessions was 48.5 minutes. This is approximately two times longer that the average session duration of students (approximately 25.5 minutes) reported in [17]. However, 48.5 minutes observation incorporated machine generated traffic from non-interactive services. After filtering the machine generated subsequences it has been discovered that the average delays between subsequences lasted approximately 6 minutes 28 seconds. This is when the knowledge workers (or services) were exposed to the passive interactions (e.g. reading a document, or filling up forms). Although in many cases the displayed material required longer processing time by humans, they avoided the excess information and proceeded to their following browsing subtask. Users' attention span was significantly shortened. It is therefore advisable to limit the passive interactions to approximately seven minutes.

The presented findings have significant implications for business process design and implementation via web services. Numerous enterprise and organizational intranet portals, e-government and e-learning portals, and also customer related interactive web services generally do not adhere to the mentioned principles. Resulting effect is low usability of intranet portals, inefficiency of e-government and e-learning web initiatives/programs, and customer dissatisfaction—leading to potential economic loss. In many instances the effective adjustments to the existing engineering and design of web services may be minimal; whether in terms of cost and/or effort. However, there are also numerous future challenges for practitioners and managers.

The managerial and design challenges lie in re-engineering the conventional business processes into the form matching the natural human browsing elements and characteristics. The identified principles should serve as an essential guideline. In addition to the indispensable general guidelines the extensive personalization technology should be employed in interactive web services. Effective personalization may vary from simple preference and profiling management to intelligent adaptive personalization technologies. Furthermore, the design should account for efficient interruption management and automated assistance services.

5 Conclusions

The current study elucidates the essential elements of human interactions in electronic environments in relation to the web services. It presents several pertinent findings for managers, designers, and developers of web services, as well as general practitioners. The results are drawn from the exploratory analysis of knowledge worker interactions on a large corporate intranet portal.

It has been observed that the knowledge workers utilize relatively small spectrum of available services. The richness and diversity of web services does not necessarily imply higher usability of the portal. Interactive web services should limit the number of required human interactions to approximately five, on average. Active interaction dynamics should be within the range of seconds. Passive human-service interactions should not necessitate more than seven minutes of human attention, on average. Complex business processes should be segmented into sub-processes reflecting these findings. There should be at most three interactive sub-process segments. The web services should feature flexible personalization and automated assistance management, in order to improve interactive efficiency and usability.

Acknowledgment

The authors would like to thank Tsukuba Advanced Computing Center (TACC) for providing raw web log data.

References

1. Zhang, J., Chang, C.K., Zhang, L.J., Hung, P.C.K.: Toward a service-oriented development through a case study. IEEE Transactions on Systems, Man and Cybernetics, Part A 37, 955–969 (2007)
2. Papazoglou, M.P., Georgakopoulos, D.: Service-oriented computing: Introduction. Communications of the ACM 46, 24–28 (2003)
3. Elfatatry, A.: Dealing with change: Components versus services. Communications of the ACM 50, 35–39 (2007)
4. Treiber, M., Dustdar, S.: Active web service registries. Internet Computing 11, 66–71 (2007)

5. Peltz, C.: Web services orchestration and choreography. Computer 36, 46–52 (2003)
6. Guerin, R.A., Pla, V.: Aggregation and conformance in differentiated service networks: a case study. ACM SIGCOMM Computer Communication Review 31, 21–32 (2001)
7. Xiong, L., Chitti, S., Liu, L.: Preserving data privacy in outsourcing data aggregation services. ACM Transactions on Internet Technology 7(17) (2007)
8. Carminati, B., Ferrari, E., Hung, P.C.K.: Exploring privacy issues in web services discovery/agencies. IEEE Security and Privacy Magazine 3, 14–21 (2005)
9. Zaha, J.M., Dumas, M., ter Hofstede, A., Barros, A., Decker, G.: Service interaction modeling: Bridging global and local views. In: Proceedings of the International Conference on Enterprise Distributed Object Computing, Hong Kong, pp. 45–55 (2006)
10. Vinoski, S.: Web services interaction models: Current practice. Internet Computing 6, 89–91 (2002)
11. Barabasi, A.-L.: The origin of bursts and heavy tails in human dynamics. Nature 435, 207–211 (2005)
12. Park, Y.-H., Fader, P.S.: Modeling browsing behavior at multiple websites. Marketing Science 23, 280–303 (2004)
13. Moe, W.W.: Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. Journal of Consumer Psychology 13, 29–39 (2003)
14. Géczy, P., Akaho, S., Izumi, N., Hasida, K.: Knowledge worker intranet behaviour and usability. Int. J. Business Intelligence and Data Mining 2, 447–470 (2007)
15. Géczy, P., Akaho, S., Izumi, N., Hasida, K.: Usability analysis framework based on behavioral segmentation. In: Psaila, G., Wagner, R. (eds.) Electronic Commerce and Web Technologies, pp. 35–45. Springer, Heidelberg (2007)
16. Gluck, J., Bunt, A., McGrenere, J.: Matching attentional draw with utility in interruption. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, San Jose, California, USA, pp. 41–50 (2007)
17. Catledge, L., Pitkow, J.: Characterizing browsing strategies in the world wide web. Computer Networks and ISDN Systems 27, 1065–1073 (1995)

Specifying Separation of Duty Constraints in BPEL4People Processes

Jan Mendling[1], Karsten Ploesser[2], and Mark Strembeck[3]

[1] BPM Cluster, Faculty of IT, Queensland University of Technology
126 Margaret Street, Brisbane QLD 4000, Australia
j.mendling@qut.edu.au
[2] SAP Research, Brisbane QLD 4000, Australia
karsten.ploesser@sap.com
[3] Vienna University of Economics and Business Administration
Institute of Information Systems, New Media Lab, Austria
mark.strembeck@wu-wien.ac.at

Abstract. Security issues have to be carefully considered for information systems that support the business processes of an organization, in particular, when these systems build on open interfaces such as web services. In this paper, we examine the new BPEL extension *BPEL4People* from an access control perspective. In particular, we discuss the importance of "separation of duty" constraints and identify options to specify such constraints in BPEL4People processes. Moreover, we identify and discuss shortcomings of the BPEL4People specifications that complicate and/or impede separation of duty enforcement. In addition, we suggest solutions which can be introduced into future versions of BPEL4People to mitigate those shortcomings.

1 Introduction

The standardization of business process management and workflow technology has been discussed for more than ten years, and several standardization bodies have proposed specifications for different aspects of business process management (see, e.g., [16]). Since 2003, the Organization for the Advancement of Structured Information Standards (OASIS) has driven the specification of the Business Process Execution Language for Web Services (BPEL) [4] which has become an important standard in this area thanks to its extensive support by major software vendors.

Originally, the BPEL specification was lacking a generic concept for activities that are performed by human agents. Due to this missing feature, BPEL could hardly be used as a platform-independent format for describing and exchanging human workflows. For this reason, major software vendors have been working on the specification of human activities in BPEL based on its extension mechanism. As a result, two complementary draft specifications were recently released: BPEL4People [2] and WS-HumanTask [1]. WS-HumanTask defines the lifecycle and the generic roles associated with a particular task, and BPEL4People defines

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 273–284, 2008.

how such tasks can be integrated in a BPEL process. In this paper, we refer to both documents together as the BPEL4People specifications, or just B4P/HT.

While the introduction of B4P/HT is a valuable extension to BPEL, it raises several questions from an access control perspective. When performing an IT-supported workflow, human users and proactive/autonomous software agents have to fulfill certain tasks to process the workflow. Each action in a workflow (like changing a document or sending a message) is typically associated with a certain access operation (e.g. to a document or a messaging service). Thus, an active entity participating in a workflow (be it a human user or a software agent) must be authorized to perform the actions that are needed to complete its tasks. The business processes of an organization are therefore an ideal source to define a tailored set of *access control policies* (also referred to as *authorization policies*) for this organization, respectively their information system (see [18,25]).

Access control deals with the elicitation, specification, maintenance, and enforcement of authorization policies in software-based systems [15,22]. In role-based access control (RBAC) [10], roles are used to model different job-positions and scopes of duty within a particular organization and/or within an information system. These roles are equipped with the permissions that are needed to perform their respective tasks. Human users and/or other active entities (subjects) are assigned to roles according to their duties, respectively their work profile (see [18,24]). The descriptions of roles tend to change significantly slower than the assignment of individuals to these roles. Thus, establishing roles as an abstraction mechanism for subjects facilitates the administration of the access control policies. Moreover, the advantages of RBAC on the modeling and technical level directly translate into lower maintenance costs [11].

A particular access control specification is said to be *safe* iff no subject can obtain an "unauthorized" right. However, since the verification of the safety property for general access control models like RBAC is not decidable [13], *constraints* are often used to enforce the safety property via explicit modeling-level artifacts. *Separation of duty constraints* enforce conflict of interest policies (see, e.g., [9,23]). Conflict of interest arises as a result of the simultaneous assignment of two mutual exclusive permissions or roles to the same subject. *Mutual exclusive* roles, or permissions result from the division of powerful rights or responsibilities to prevent fraud and abuse. An example is the common practice to separate the "controller" role and the "chief buyer" role in medium-sized and large companies. Two mutual exclusive roles are not allowed to be assigned to the same subject, and two mutual exclusive permissions must not be assigned to the same role or the same subject. Moreover, in workflow environments it is of central significance that separation of duty constraints can also be defined and enforced on the level of tasks (see, e.g., [7]).

There has been some work on access control for BPEL processes (see, e.g., [5,14,17]). However, an analysis of the B4P/HT specification from a separation of duty perspective is missing so far. In this paper, we thus address the enforceability of separation of duty constraints in B4P/HT. This includes questions like: How can we express in B4P/HT that a certain individual is not allowed to do both the “check loan application” task and the “decide about loan acceptance” task for one particular loan application? And if B4P/HT provides means to express such separation of duty constraints, are there ways to circumvent them?

The remainder of this paper is structured as follows. In Section 2 we introduce the concepts of B4P/HT which are important for the purposes of this paper. Section 3 then gives an overview of role-based and task based access control, and discusses the relevance of separation of duty constraints. Subsequently, Section 4 provides an analysis of B4P/HT from the separation of duty perspective. In particular, we discuss how specific aspects of separation of duty constraints can be enforced in B4P/HT, identify corresponding shortcomings of the B4P/HT specifications, and suggest solutions which can be introduced into future versions to mitigate those shortcomings. Next, Section 5 discusses related work before Section 6 concludes the paper.

2 Preliminaries on B4P and WS-HT

In this section we give an overview of BPEL4People and WS-HumanTask (B4P/HT). For this introduction there is no detailed knowledge of BPEL required. In essence, a BPEL Process defines a set of activities and their control flow. BPEL 2.0 [4] offers a so-called “extension activity” which is used by B4P/HT for adding human tasks to BPEL. B4P/HT defines three generic process roles: the process initiator, process stakeholder, and business administrator. While in most cases the process initiator will be determined by the BPEL engine at runtime, the other two generic roles are populated by evaluating a people query. We use the term “people query” to abstract from the three different options provided by B4P/HT of assigning people to roles, i.e. using the so-called logical people group which is an alias to a query, using literals, or using expressions (see [1,2]). The specification, however, does not mandate a concrete people query language, which leaves room for vendor-specific implementations. The BPEL process related concepts of B4P/HT are depicted on the left-hand side of Figure 1.

The right-hand side of this figure shows the task concept, its generic roles, the task lifecycle states, and lifecycle transitions. These concepts are defined in the WS-HumanTask specification. A single task has different generic roles assigned to it. Similar to the overall process there are task initiators, task stakeholders and business administrators. Beyond that, a task has a set of potential owners, one actual owner, and a list of excluded owners. These three generic roles define in essence who *may* legally be assigned to a task and who *is* actually assigned to a task. The people that populate these different task generic roles can influence the lifecycle of a task at predefined stages.

The typical sequence of B4P/HT task states is from *created* to *ready* to *reserved* to *in-progress* to *completed* [1, p.37]. Depending on the task state and on their assignment to one of the generic task roles, people can control the task lifecycle via predefined transitions. Several of these transitions are only allowed for members of the *potential owner* role, including to claim and to start, to suspend and to resume, as well as to delegate and to forward a task. Both the latter

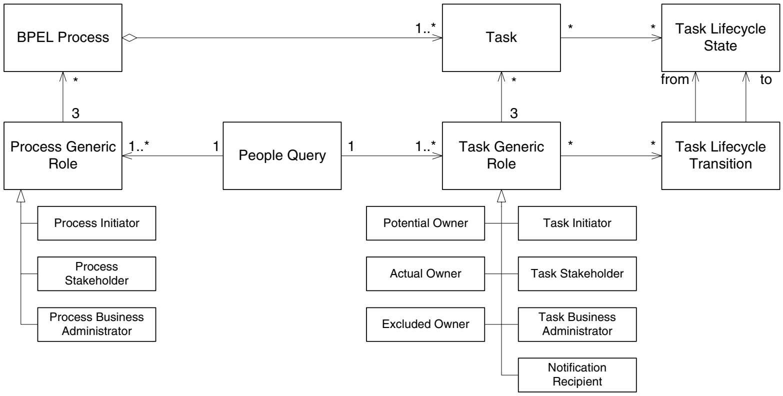


Fig. 1. Class diagram: BPEL4People

(forward and delegate) are interesting from a security perspective. When a task is delegated, the delegatee becomes the actual owner of the task and is added to the potential owners list if she was not before (see also Section 4). The set of potential delegatees can be limited for each task to "anybody", "nobody", "potential owners", or to a list of predefined people ("others"). The forwarding mechanism works similar with the only difference that it cannot be restricted and that the forwardee replaces the forwarder in the potential owners list. Most of these transitions are also allowed for members of the *actual owner* role. Beyond that, an actual owner can stop, release, complete, signal failure, and skip a task.

The *business administrator* is the most powerful role since all state transitions are accessible to it. The business administrator is important if the list of potential owners is evaluated to the empty set. In this case, he nominates one person to execute the task. Beyond that, B4P/HT allows to define that certain people may not perform a particular task via the generic role *excluded owners*. Users being assigned to the "excluded owners" role are thus excluded from the potential owners list, and cannot progress the task in its lifecycle.

Furthermore, B4P/HT standardizes methods to access the assignment history of a particular workflow case. These methods can be used to query the members of a generic role for a particular task. For instance, the *getActualOwner(X)* function returns the actual owner of task X. This way, one can retrieve role members for previously executed tasks, for instance, to exclude them from other tasks. We will elaborate on this feature later when we discuss how B4P/HT can be tailored to support task-based separation of duty constraints.

3 Role-Based and Task-Based Access Control

In order to allow for an (automated) enforcement of authorization policies, the high-level control objectives specified for a system need to be mapped to the structures provided by an access control model. An *access control model* provides an abstract framework for the definition of authorization policy rules. It defines how essential access control elements (like subjects, operations, objects) could be interrelated. In addition, it may specify invariants which must be met by each real-world implementation of this model (e.g. to enable control of information flows) or predetermine administrative procedures.

In recent years, role-based access control (RBAC) [10] – together with diverse extensions and variants – has evolved into a de facto standard for access control in both research and industry. One of the advantages of RBAC is being a general access control model. This means that a sophisticated RBAC-service may be configured to emulate many different access control models, including discretionary and mandatory access control models (see [19]).

A central idea in RBAC is to support constraints on almost all parts of an RBAC model (e.g. permissions, roles, or assignment relations) to achieve high flexibility. Static and dynamic separation of duty constraints (see [8]) are two of the most common types of RBAC constraints (see, e.g., [3]). Separation of duty (SOD) constraints can be subdivided in static separation of duty (SSD) constraints and dynamic separation of duty (DSD) constraints. *Static separation of duty* constraints specify that two mutual exclusive roles (or permissions) must *never* be *assigned* to the same subject *simultaneously*. *Dynamic separation of duties* constraints define that two mutual exclusive roles (or permissions) must *never* be *activated* by the same subject *simultaneously*. This means that two dynamically mutual exclusive roles may be assigned to the same subject. The corresponding subject, however, is only allowed to activate at most one of its dynamically mutual exclusive roles (permissions) at the same time.

Various contributions concerning access control in collaborative environments exist, esp. for groupware and workflow systems. For example, Thomas and Sandhu introduced TBAC [29], a family of models that support the specification of active security models. In TBAC, permissions are actively (de)activated according to the current task/process-state. In [6], Bertino et al. present a well-elaborated language and algorithms to express and enforce constraints which ensure that all tasks within a workflow are performed by predefined users/roles. In [12], Georgiadis et al. introduce the Context-based Team Access Control model (C-TMAC) as an extension of the TMAC approach presented by Thomas [28]. Here, a team is defined as a group of users acting in different roles with the objective of corporately completing a certain task, e.g. a group of physicians and nurses attending a patient. Thus, in C-TMAC the team concept is used to associate users with contexts, like roles are used to associate users with permissions. One similarity for all of these approaches is that they facilitate the usage of some context information, e.g. the execution history of individuals/roles and the current task, to make assignment, activation, or authorization decisions.

In this paper, we are especially interested in the specification and enforcement of task-based separation of duty constraints in the B4P/HT context. Here, a *task-based separation of duty constraint* is a separation of duty constraint that considers task order and task history in a particular process instance to decide if a certain subject or role is allowed to perform a certain task (see also [7,30]).

Again, task-based SOD constraints can be static or dynamic. A *static task-based SOD constraint* defines that two statically mutual exclusive tasks must never be assigned to the same role and must never be performed by the same subject. This constraint is global with respect to all process instances in the corresponding information system. For example, a company may choose to define two tasks "Order supplies" and "Approve payment" as statically mutual exclusive to prevent fraud and abuse.

In contrast, a *dynamic task-based SOD constraint* refers to individual process instances and defines that two dynamically mutual exclusive tasks must never be performed by the same subject in the same process instance. In other words: two dynamically mutual exclusive tasks can be assigned to the same role. However, to complete a process instance which includes two mutual exclusive tasks, one needs at least two different subjects (i.e. two individuals owning the respective role). This means, although a subject might possess a role which includes all permissions to perform two dynamically mutual exclusive tasks, a dynamic task-based SOD constraint can enforce that the same subject does not perform both tasks in the same process instance. For example, a bank may assign two tasks "Check credit worthiness" and "Approve credit application" to the "Bank clerk" role and define a dynamic task-based SOD constraint on these tasks. Each subject owning the bank clerk role may then perform both tasks. Nevertheless, because of the dynamic SOD constraint on these tasks, we always need at least two bank clerks to complete a "Credit application" process.

Figure 2 shows the essential relationships of subjects (users), roles, permissions, tasks, and processes that are important for the purposes of this paper.

In general, static as well as dynamic SOD constraints can be defined on the level of roles, permissions, and tasks – resulting in a total of six different types of SOD constraints. Moreover, SOD constraints are subject to inheritance (see, e.g.,

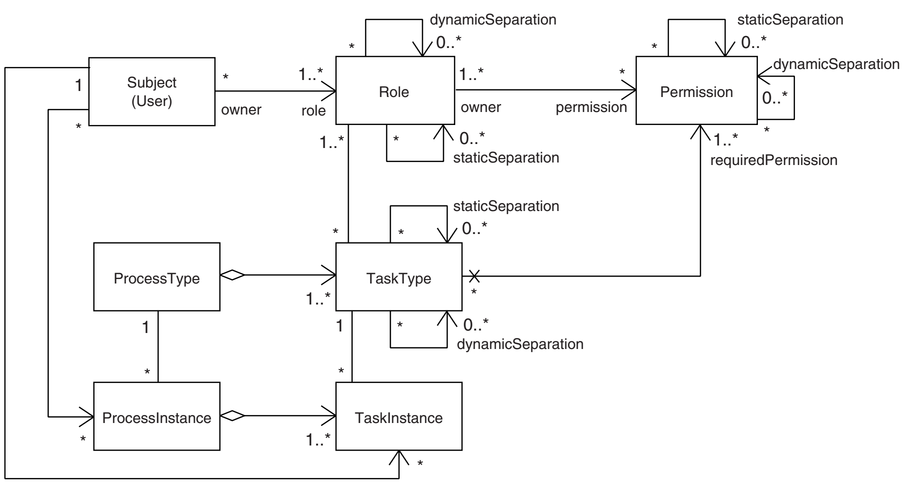


Fig. 2. Class diagram: task and role-based access control

[9,23]). For example, if one of two statically mutual exclusive tasks is assigned to a certain role, the other task must be assigned to an other role. Subsequently, the corresponding roles are statically mutual exclusive because they inherit the corresponding SOD constraints from their assigned tasks. Moreover, two tasks are mutual exclusive if one needs two mutual exclusive permissions to complete both tasks. In this case, the SOD constraint between two tasks is inherited from the permissions that are needed to perform these tasks (see also Figure 2).

4 BPEL4People Support for SOD Constraints

In this section, we discuss in how far the six types of separation of duty constraints (see Section 3) can be implemented in B4P/HT. First, Section 4.1 presents strategies for adding SOD support to a B4P/HT process. Subsequently, Section 4.2 discusses access control enforcement issues in B4P/HT and proposes solutions.

4.1 Strategies for Implementing SOD in B4P/HT

WS-B4P/HT does not advocate a specific mechanism to implement access control. More specifically, the specification states that a "mechanism determining how an implementation evaluates people assignments" is out of scope. Given this design choice there are basically two ways to implement access control within B4P/HT: firstly, via the enforcement of access control policies in people queries, and secondly, by enforcing access control policies in the people assignment section of the respective tasks. At the time of writing, there is no standardized people query language available to be used with WS-B4P/HT. Accordingly, no assumptions can be made about access control mechanisms in such a language. As a consequence of that, we focus on the second approach[1].

Role-based constraints play an important part in the specification of access control. However, even though B4P/HT utilizes people queries as a role-like concept, it does not provide means to define direct relations between roles such as inheritance relations or separation of duty relations. More precisely, B4P/HT states that "the structure of the data in the people directory is [. . .] out of scope" (see [1,2]). Below, we now discuss if and how the different types of separation of duty constraints can be supported based on the options provided through the current B4P/HT specifications:

Permission-based SOD: Since B4P/HT does not support a notion of permission, corresponding SOD constraints need to be captured in people queries or in the organization model outside B4P/HT.

[1] Note that a third option is the extension of B4P/HT with explicit (implementation independent) concepts to enable the direct and explicit integration of access control relevant information in B4P/HT task definitions. Our analysis provides the foundation for considering suitable extensions.

Static role-based SOD: In essence, static role-based separation of duty demands that two mutually-exclusive roles must not be assigned to the same subject. This requirement can be partially translated into the people query concept of B4P/HT. When two people queries pq_1 and pq_2 should be mutually exclusive, this can be enforced via the generic role assignment of a task in the following way: tasks whose potential owners are populated with pq_1 get pq_2 assigned to the excluded owners (see Section 2). As a consequence, there will be no task in the process for which pq_1 and pq_2 can yield the same subject.

Dynamic role-based SOD: Dynamic role-based SOD demands that two mutual exclusive roles can never be activated by the same subject. We again consider two people queries pq_1 and pq_2 for this case. This requirement can be enforced by adding each actual owner ao_1 and ao_2 of all tasks that use pq_1 and pq_2 to get their potential owners as excluded owners. This means, if task t_1 has the extension of pq_1 as its potential owners, then it must exclude all actual owners of tasks that have pq_2 as potential owners. The actual owners can be retrieved using the getActualOwner method of B4P/HT.

Static task-based SOD: Static task-based separation of duty demands that two mutual exclusive tasks must not be assigned to the same role. While B4P/HT does not directly offer a user-configurable role concept, this constraint can be enforced using two people queries pq_1 and pq_2 which populate the potential owner roles po_1 and po_2 of the mutual exclusive tasks t_1 and t_2. By cross-assigning the people queries to the excluded owners of the respective other tasks, the two sets of people who can actually execute these tasks are disjoint.

Dynamic task-based SOD: Dynamic task-based separation of duty demands that mutual exclusive tasks must not be executed by the same subject in the same process instance. This concept can be directly represented in B4P/HT by using the getActualOwner method, and assigning it to the excluded owner role of the respective other task.

4.2 Access Control Enforcement Issues in B4P/HT

In the previous section, we discussed how role-based and task-based separation of duty constraints can be captured in B4P/HT. Yet, one needs to be aware of some limitations of this approach, in particular, regarding task delegation and forwarding, the getActualOwner method, and the role of the business administrator:

– The proposal that we make in the previous section can only partially avoid delegation and forwarding of B4P/HT tasks. While the delegation parameters of a task can be explicitly set, it is, according to the current B4P/HT specification, not possible to switch off the forwarding mechanism. Albeit the specification states that "excluded owners are implicitly removed from the set of potential owners", it is unclear when and how this requirement is enforced. Thus, a potential owner, who is assigned to a task instance in accordance with the SOD constraints, can potentially forward this task instance to an unauthorised user. In general, it is advisable to disallow forwarding in

certain scenarios to prevent forwarding operations which would violate the active set of SOD constraints for a task. Therefore, we recommend clarifying the enforcement mechanism or providing means to explicitly switch off forwarding in a future version of B4P/HT.
- Moreover, we encountered an issue with evaluating the getActualOwner method if dynamic task-based SOD constraints are defined for two concurrent tasks. In this case, a task can be started at a point in time when there is not yet an actual owner determined for the other (mutual exclusive) task. Accordingly, the excluded owner lists of both (mutual exclusive) task instances are populated with the empty set (see also Sections 2 and 4.1), and the corresponding dynamic separation of duty constraint is not enforced. Furthermore, at this stage the B4P/HT specification only allows determining the actual owner when the task reaches a final state. From an access control perspective both options – current owner as well as owner history of a task instance – are important and are necessary to make certain access control decisions. Moreover, the current actual owner needs to be known, e.g. in case of escalation. Therefore, it seems to be a good choice to include two different methods in a future version of the specification which can provide these information.
- Finally, the user acting in the business administrator role can override almost all restrictions of a B4P/HT process. While this is critical regarding the level of trust that a subject fulfilling this role would deserve, there is another issue regarding awareness of SOD constraints. A B4P/HT-compliant implementation should provide information about possible SOD violations to the business administrator when he forwards, delegates, or nominates somebody for a task. This way, he should be able to avoid assignments that contradict SOD rules.

Altogether, the result is that B4P/HT offers mechanisms to capture different aspects of SOD constraints. Still, there are some weaknesses regarding enforcement that should be fixed in future versions of the specification.

5 Related Work

While this paper provides the first analysis of the B4P/HT specification from a separation of duty perspective, several approaches exist to extend BPEL or closely related workflow notations with access control means: In [17] we presented an approach to extract RBAC models from BPEL4WS processes. Bertino et al. [5] introduce RBAC-WS-BPEL, a language for authorization policies for business processes defined in BPEL. Furthermore, they introduce the business process constraint language (BPCL) – BPCL is defined as an XML schema and allows for the specification of authorization constraints for BPEL processes, such as separation of duty or binding of duty constraints. Wolter and Schaad [30] extend the business process modeling notation (BPMN) with a means to model task-based authorization constraints. In particular, they focus on separation of duty constraints to model conflicting roles and/or conflicting tasks.

In [27] Thomas et al. propose several BPEL extensions to support user tasks and to define access control requirements of these tasks. They suggest a software component called "people activity manager" (which is, however, unrelated to the B4P/HT specifications [1,2]) that makes access control decisions for user tasks. We complement this work by analyzing the expressiveness and weaknesses of B4P. In this regard we extend previous work on a workflow resource pattern evaluation of B4P that only briefly touches access control issues [20,21].

6 Conclusion

In this paper, we analyzed options how the current B4P/HT specifications can support different types of separation of duty constraints. In particular, we discussed how people queries, in conjunction with generic B4P/HT roles, can be used to enforce several aspects of SOD constraints. Moreover, we identified limitations in the current B4P/HT specification and propose solutions to address these limitations in future versions of B4P/HT.

In particular, we suggest providing either a clarification of the enforcement mechanism for excluded owners or a mechanism to explicitly switch off task forwarding in a future version of B4P/HT – this is to prevent users from circumventing SOD constraints via B4P/HT's forward mechanism. Second, the B4P/HT specification only allows to determine the actual owner of a task when the task reaches its final state. However, from an access control perspective it can be important to determine the current owner as well as the owner history of a certain task instance. Thus, we propose to include two different methods in a future version of the B4P/HT specification – one method that returns the current task owner only, and an other method that returns the owner history of a task. Third, because the B4P/HT business administrator role could override almost all restrictions defined for a B4P/HT process, we suggest that B4P/HT-compliant implementations make a user acting as business administrator aware of potential SOD violations when he intends to forward, delegate, or nominate somebody for a task. This way, a user acting in the business administrator role can avoid assignments that contradict SOD rules. An other option could be a restriction of the B4P/HT business administrator role, which, however, may be an undesired option.

In addition to our suggestions summarized above, a native B4P/HT extension would be beneficial to provide explicit (implementation independent) concepts that enable the direct and explicit integration of access control relevant information in B4P/HT task definitions. In our future work, we therefore aim to define such native extensions. Furthermore, when using separation of duty constraints in workflow environments, it is important to enable consistency checks on tasks and corresponding constraint specifications, to guarantee that the constraints properly control task execution and user-to-task assignment without preventing the processes from being completed (see, e.g., [26,30]) Thus, we also intend to provide such features for B4P/HT environments in our future work. Finally, there is a need for an efficient and consistent approach to specify access control information in B4P/HT. In future research, we will investigate the suitability of a model-driven approach to generate B4P/HT based on the rules we identify in this paper to express separation of duty.

References

1. Agrawal, A., Amend, M., Das, M., Keller, C., Kloppmann, M., König, D., Leymann, F., Müller, R., Pfau, G., Ploesser, K., Rangaswamy, R., Rickayzen, A., Rowley, M., Schmidt, P., Trickovic, I., Yiu, A., Zeller, M.: Web services human task (WS-HumanTask), version 1.0 (2007)
2. Agrawal, A., Amend, M., Das, M., Keller, C., Kloppmann, M., König, D., Leymann, F., Müller, R., Pfau, G., Ploesser, K., Rangaswamy, R., Rickayzen, A., Rowley, M., Schmidt, P., Trickovic, I., Yiu, A., Zeller, M.: WS-BPEL extension for people (BPEL4People), version 1.0 (2007)
3. Ahn, G.J., Sandhu, R.: Role-based Authorization Constraints Specification. ACM Trans. on Information and System Security (TISSEC) 3(4) (November 2000)
4. Alves, A., Arkin, A., Askary, S., Barreto, C., Bloch, B., Curbera, F., Ford, M., Goland, Y., Guizar, A., Kartha, N., Liu, C.K., Khalaf, R., Koenig, D., Marin, M., Mehta, V., Thatte, S., van der Rijn, D., Yendluri, P., Yiu, A.: Web Services Business Process Execution Language - Version 2.0. OASIS (January 2007)
5. Bertino, E., Crampton, J., Paci, F.: Access Control and Authorization Constraints for WS-BPEL. In: Proc. of the IEEE International Conference on Web Services (ICWS) (September 2006)
6. Bertino, E., Ferrari, E., Atluri, V.: The Specification and Enforcement of Authorization Constraints in Workflow Management Systems. ACM Transactions on Information and System Security (TISSEC) 2(1) (February 1999)
7. Botha, R.A., Eloff, J.H.P.: Separation of duties for access control enforcement in workflow environments. IBM Systems Journal 40(3) (2001)
8. Clark, D.D., Wilson, D.R.: A Comparison of Commercial and Military Computer Security Policies. In: Proc. of the IEEE Symposium on Security and Privacy (1987)
9. Ferraiolo, D.F., Barkley, J.F., Kuhn, D.R.: A Role-Based Access Control Model and Reference Implementation within a Corporate Intranet. ACM Transactions on Information and System Security (TISSEC) 2(1) (February 1999)
10. Ferraiolo, D.F., Sandhu, R., Gavrila, S., Kuhn, D.R., Chandramouli, R.: Proposed NIST Standard for Role-Based Access Control. ACM Transactions on Information and System Security (TISSEC) 4(3) (August 2001)
11. Gallaher, M.P., O'Connor, A.C., Kropp, B.: The Economic Impact of Role-Based Access Control. National Institute of Standards & Technology (NIST), Planning Report 02-1 (March 2002)
12. Georgiadis, C.K., Mavridis, I., Pangalos, G., Thomas, R.K.: Flexible Team-Based Access Control Using Contexts. In: Proc. of the 6th ACM Symposium on Access Control Models and Technologies (SACMAT) (May 2001)
13. Harrison, M.A., Ruzzo, W.L., Ullman, J.D.: Protection in Operating Systems. Communications of the ACM 19(8) (August 1976)
14. Koshutanski, H., Massacci, F.: Interactive access control for web services. In: Deswarte, Y., Cuppens, F., Jajodia, S., Wang, L. (eds.) IFIP 18th WorldComputer Congress, TC11 19th Int. Information Security Conference, pp. 151–166 (2004)
15. Landwehr, C.E.: Formal Models for Computer Security. ACM Computing Surveys 13(3) (September 1981)

16. Mendling, J., zur Muehlen, M., Price, A.: Process Aware Information Systems: Bridging People and Software Through Process Technology. In: Standards for Workflow Definition and Execution, pp. 281–316. Wiley Publishing, Chichester (2005)
17. Mendling, J., Strembeck, M., Stermsek, G., Neumann, G.: An Approach to Extract RBAC Models from BPEL4WS Processes. In: Proc. of the 13th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises (WETICE) (June 2004)
18. Neumann, G., Strembeck, M.: A Scenario-driven Role Engineering Process for Functional RBAC Roles. In: Proc. of 7th ACM Symposium on Access Control Models and Technologies (SACMAT) (June 2002)
19. Osborn, S., Sandhu, R., Munawer, Q.: Configuring Role-Based Access Control to Enforce Mandatory and Discretionary Access Control Policies. ACM Transactions on Information and System Security (TISSEC) 3(2) (February 2000)
20. Russell, N., van der Aalst, W.M.P.: Evaluation of the BPEL4People and WS-HumanTask Extensions to WS-BPEL 2.0 using the Workflow Resource Patterns. BPM Center Report BPM-07-10, BPMcenter.org (2007)
21. Russell, N., van der Aalst, W.M.P., ter Hofstede, A.H.M., Edmond, D.: Workflow resource patterns: Identification, representation and tool support. In: Pastor, Ó., Falcão e Cunha, J. (eds.) CAiSE 2005. LNCS, vol. 3520, pp. 216–232. Springer, Heidelberg (2005)
22. Sandhu, R.S., Samarati, P.: Access control: Principles and practice. IEEE Communications 32(9) (September 1994)
23. Strembeck, M.: Conflict Checking of Separation of Duty Constraints in RBAC - Implementation Experiences. In: Proc. of the Conference on Software Engineering (SE 2004) (February 2004)
24. Strembeck, M.: A Role Engineering Tool for Role-Based Access Control. In: Proc. of the 3rd Symposium on Requirements Engineering for Information Security (SREIS) (August 2005)
25. Strembeck, M.: Embedding Policy Rules for Software-Based Systems in a Requirements Context. In: Proc. of the IEEE International Workshop on Policies for Distributed Systems and Networks (POLICY) (June 2005)
26. Tan, K., Crampton, J., Gunter, C.A.: The Consistency of Task-Based Authorization Constraints in Workflow Systems. In: Proc. of the IEEE Workshop on Computer Security Foundations (CSFW) (June 2004)
27. Thomas, J., Paci, F., Bertino, E., Eugster, P.: User Tasks and Access Control over Web Services. In: Proc. of the IEEE International Conference on Web Services (ICWS) (July 2007)
28. Thomas, R.K.: Team-based Access Control (TMAC): A Primitive for Applying Role-based Access Controls in Collaborative Environments. In: Proc. of the ACM Workshop on Role Based Access Control (1997)
29. Thomas, R.K., Sandhu, R.S.: Task-based authorization controls (TBAC): A family of models for active and enterprise-oriented authorization management. In: Proc. of the IFIP WG11.3 Conference on Database Security (August 1997)
30. Wolter, C., Schaad, A.: Modeling of Task-Based Authorization Constraints in BPMN. In: Alonso, G., Dadam, P., Rosemann, M. (eds.) BPM 2007. LNCS, vol. 4714, pp. 64–79. Springer, Heidelberg (2007)

Web Service Discovery in the FUSION Semantic Registry

Dimitrios Kourtesis and Iraklis Paraskakis

South East European Research Centre (SEERC),
Mitropoleos 17, 54624 Thessaloniki, Greece
{dkourtesis,iparaskakis}@seerc.org

Abstract. The UDDI specification was developed as an attempt to address the key challenge of effective Web service discovery and has become a widely adopted standard. However, the text-based indexing and search mechanism that UDDI registries offer does not suffice for expressing unambiguous and semantically rich representations of service capabilities, and cannot support the logic-based inference capacity required for facilitating automated service matchmaking. This paper provides an overview of the approach put forward in the FUSION project for overcoming this important limitation. Our solution combines SAWSDL-based service descriptions with service capability profiling based on OWL-DL, and automated matchmaking through DL reasoning in a semantically extended UDDI registry.

Keywords: Semantic Web Services, Web Service Discovery, Universal Description Discovery and Integration (UDDI), Semantic Annotations for WSDL (SAWSDL), Enterprise Interoperability.

1 Introduction

The Service Oriented Architecture (SOA) paradigm and its manifestation in the form of the Web services technology stack promise to become prime enablers for business agility in the modern enterprise by alleviating many of the barriers that stand on the path to achieving Enterprise Application Integration (EAI). Integrating a set of service-oriented business applications necessitates the assembly of services exposed by the individual business applications into new service compositions. This in turn requires discovering services that are suitable for performing each of the key tasks that a business process workflow comprises. Notably, in a fully SOA-enabled business application ecosystem with hundreds of deployed Web services, the task of searching and identifying the ones that are most appropriate for a certain type of use can become rather demanding.

This was the motivation behind the development of the Universal Description, Discovery and Integration (UDDI) specification [1] as a standardised way to catalogue and discover reusable services. UDDI registries however lack the means for supporting automated service discovery [2], [3], [4]. The reason is that indexing and retrieval in UDDI is not based on unambiguous, semantically rich representations of

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 285–296, 2008.

Web service capabilities but on unstructured textual descriptions and categorisations that are retrievable through keyword-based search. Keyword-based annotation and search techniques cannot facilitate automated discovery since they do not provide any way of differentiating among (i) services that have identical naming but perform totally unrelated operations and (ii) services that have totally different names but offer equivalent functionality. To illustrate this problem through real-world examples, consider the case of two Web services that share “Address Validation” as their name but offer different functionality: the first one[1] validates postal addresses in the United States, while the second one[2] checks the validity of email addresses. Furthermore, consider the case of a service categorised by the name “UK Location”[3], able to check the validity of United Kingdom postal addresses, and another service categorised by the name “Global Address Verification”[4] which can still be of use for the exact same purpose, despite its apparently counter-intuitive name.

To overcome the problem of ambiguity that hinders automated service discovery we need to describe service characteristics in a formal, machine-understandable manner that is amenable to processing within semantically-enhanced UDDI registries. The aim of this paper is to present the approach adopted in project FUSION and the open source FUSION Semantic Registry[5] towards this direction, improving and elaborating on the preliminary work presented in [5]. FUSION is an EU-funded research project[6] aiming to promote efficient business process integration within and across enterprises, by offering a semantics-based solution to achieving interoperability among service-oriented business applications. The project aims at delivering a complete reference framework and a methodology for semantics-based Enterprise Application Integration (EAI), a reference implementation of the proposed framework, and a validation of the proposed approach through three pilot studies on intra- and inter-organisational integration. The introduction of semantics to Web service discovery is an essential requirement for realising the approach that FUSION puts forward, and encompasses: (i) describing service advertisements and service requests in a way that is formal, unambiguous and semantically precise, and (ii) realising a UDDI-based service registry that offers semantically-enhanced publication and discovery functions.

The rest of this paper is organised as follows. Section 2 introduces the requirements that FUSION puts forward for semantically describing service advertisements and service requests. Section 3 presents an overview of the FUSION Semantic Registry architecture, and provides a walkthrough on the core activities performed during service publication and service discovery. Section 4 analyses the matchmaking capabilities that the FUSION Semantic Registry supports, and its applicability for evaluating the relevance among a service advertisement and a service request at three distinct levels. Section 5 gives an overview and comparison of related work in this area, and section 6 concludes the paper with a small synopsis of the topics discussed.

---

[1] http://ws2.serviceobjects.net/av/AddressValidate.asmx?WSDL
[2] http://service.ecocoma.com/email/validate.asmx?WSDL
[3] http://www.webservicex.net/uklocation.asmx?WSDL
[4] http://ws.strikeiron.com/GlobalAddressVerification4?WSDL
[5] http://www.seerc.org/fusion/semanticregistry/
[6] http://www.fusion-strep.eu/

2 Describing Service Characteristics in FUSION

By using a semantic representation formalism to express the characteristics of Web services offered or needed, providers and requestors create definitions of service capabilities that are automatically processable through reasoning and logic-based inference. In turn, this can facilitate high-precision retrieval for services residing in a semantically-enhanced service registry, and offer a significant improvement over the capabilities of conventional UDDI registries. Evidently, the extent to which this can be achieved depends on the semantic representation formalism that is adopted for this purpose. The recent years have seen numerous Semantic Web Service frameworks being proposed and promoted for standardisation through W3C member submissions. The most prominent ones are OWL-S [6], WSMO [7], and the WSDL-S [8] specification that evolved into the W3C Recommendation of SAWSDL [9].

Although the FUSION reference framework does not prescribe the use of a specific Semantic Web Service framework, the reference implementation of the FUSION System that the Semantic Registry is part of builds on SAWSDL. In contrast to developing Web service descriptions at a high conceptual level and then linking these specifications to concrete Web service interfaces that are described in WSDL (as proposed in OWL-S and WSMO), the approach that SAWSDL puts forward is bottom-up: the WSDL documents themselves are to be enriched with annotations that capture machine processable semantics by pointing to concepts defined in externally maintained semantic models. The advantages of this approach are many-fold, but the most important one is that SAWSDL becomes agnostic to the knowledge representation formalism that one adopts. This allows service providers to annotate their services with concepts described in any modelling language, provided that these concepts are uniquely identifiable through URIs so that they can be referenced from within annotations. This promotes reusability for existing domain models and even allows SAWSDL to be used in conjunction with OWL-S or WSMO to combine the best of both worlds.

The semantic model that serves as the basis for creating, storing, and reasoning upon representations of service capabilities in the FUSION project is the FUSION Ontology [10]. Its multi-faceted structure reflects different types of concepts necessary for modelling a service: the data structures a service exchanges through messages (data semantics), the functionality categorisation of a service with regard to a taxonomy (classification semantics), and the behaviour it may expose within a complex and stateful process execution (behavioural semantics). The FUSION Ontology is encoded in OWL-DL, a Description Logics fragment of the W3C standard Web Ontology Language (OWL) that strikes a satisfactory balance between expressiveness and computational completeness [11] and facilitates decidable reasoning with the help of DL reasoning engines.

To represent the characteristics of a specific service advertisement or request in FUSION, one needs to create a Functional Profile, and define its key attributes in terms of references to the FUSION Ontology. A Functional Profile is expressed as a named OWL class that is attributed a set of three different OWL object properties:

i. `hasCategory`: associates a `FunctionalProfile` with a `TaxonomyEntity` concept from the service classification taxonomy that is part of the FUSION

Ontology, in order to represent the service's functionality categorisation. The cardinality of this property is exactly one.

ii. `hasInput`: associates a `FunctionalProfile` with an `InputDataSet` concept, in order to represent the set of data parameters that comprise the request message a service expects to receive and consume. The cardinality of this property is zero in the case of an *out-only* Message Exchange Pattern (MEP), or one, in the case of an *in-out* MEP.

iii. `hasOutput`: associates a `FunctionalProfile` with an `OutputDataSet` concept, in order to represent the set of data parameters that comprise the response message a service will produce if invoked. The cardinality of this property is zero in the case of an *in-only* MEP, or one, in the case of an *in-out* MEP.

Finally, each `InputDataSet` and `OutputDataSet` concept is associated with one or more `DataFacetEntity` concepts through a `hasDataParameter` object property, in order to represent the data parameters that comprise the message.

Depending on the perspective from which the Functional Profile is viewed, that of the provider or the requestor, we can differentiate among Advertisement Functional Profiles (AFPs) and Request Functional Profiles (RFPs). The first are created automatically by the FUSION Semantic registry at the time of service publication, while the latter are created by the service requestor at the time of discovery (or even at an earlier stage to be used as service request templates).

To allow for the construction of Advertisement Functional Profiles (AFPs), service providers need to augment the WSDL interfaces of their provided services with semantic annotations. The resulting SAWSDL interfaces must capture two elementary types of semantics: (i) the semantics of the data structures that a service exchanges through messages, and (ii) the semantics of a service's categorisation with respect to the functionality classification taxonomy. The semantics of a service's input and output data are captured by adding `modelReference` annotations to the appropriate `<xs:element>` entities under `<wsdl:types>`, while functionality categorisation semantics are captured with `modelReference` annotations on `<wsdl:portType>` entities.

3 An Overview of the FUSION Semantic Registry Architecture

There are many ways to realise a UDDI-based service registry that performs semantically-enhanced service matchmaking. A number of relevant attempts, each addressing a different set of requirements, are reviewed in section 5. The FUSION Semantic Registry architecture that is presented in this paper augments the purely syntactic search facilities that a UDDI registry can offer without requiring any modifications to the implementation of the UDDI server or the UDDI specification API, and this can be an important advantage compared to other approaches. We propose an architecture that positions a set of semantically-enabled modules externally to the UDDI server. These modules provide a specialised Web service API to the client, and are responsible for performing the necessary SAWSDL parsing, OWL ontology processing, and DL reasoning operations. Approaches based on this principle (i.e. relying on external components for specialised functionality while retaining the UDDI server implementation intact) have been also proposed in [4] and [12].

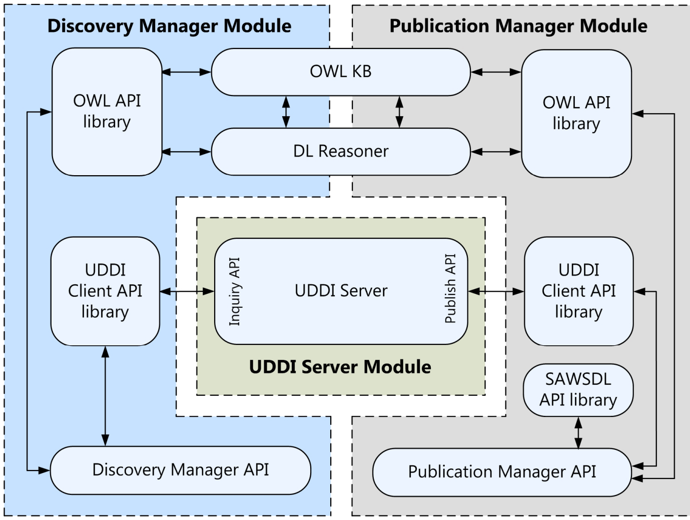


Fig. 1. FUSION Semantic Registry high-level architecture

As illustrated in Figure 1, the architecture that we propose comprises three core modules: (i) the UDDI Server Module, (ii) the Publication Manager Module, and (iii) the Discovery Manager Module. The UDDI server module is a typical server implementation of the UDDI v2 specification by OASIS [1] and a description of its functionality is beyond the scope of this paper. In the remaining of this section we focus on describing the functionality of the other core modules, and especially on the activities taking place during service publication and discovery.

3.1 Functionality of the Publication Manager Module

The Publication Manager Module provides an interface to the user for adding, removing, or updating descriptions of Web services (service advertisements), as well as adding, removing, or updating descriptions of service providers. A service provider can be a company, a business unit within an organisation, or even a specific business information system that offers some service on the network. The procedure of publishing a service advertisement comprises four phases:

i. Parsing an SAWSDL document to extract syntactic and semantic information
ii. Using the extracted semantic information to construct an Advertisement Functional Profile (AFP)
iii. Classifying the AFP in the registry's OWL Knowledge Base (KB)
iv. Mapping the extracted syntactic information and the derived semantic information to appropriate UDDI structures

The publication query that is used for initiating the publication process comprises four elements (i) the service provider ID (each service is associated with a specific service provider), (ii) a URL pointing to the SAWSDL document that contains the service description to be published (iii) an optional name for the service, and (iv) an optional free text description for the service.

The Publication Manager retrieves the SAWSDL document from the specified URL and extracts discovery-related information. Notably, the most valuable type of information to assist in discovery is not the syntactic characteristics of a service (e.g. its port and binding protocol information), but its defined categorisation and input/output messages, as already discussed in section 2. As depicted in Figure 1, this information is extracted with the help of an SAWSDL API library[7] that provides parsing and serialisation facilities. The syntactic and semantic characteristics that are extracted serve as input to a hybrid OWL-DL/UDDI indexing procedure.

Indexing begins by constructing an AFP and adding it to the registry's internal OWL Knowledge Base (KB) through the OWL API library[8] depicted in Figure 1. The Pellet DL reasoner[9] is subsequently used for performing an "eager" semantic classification of the new AFP against all known Request Functional Profiles (RFPs). The purpose of this classification procedure is to identify RFPs representing service requests that the newly added service advertisement can readily satisfy. We refer to this classification procedure as "eager" since it takes place at publication-time. In contrast, a "lazy" classification procedure would not have taken place before the actual need for matchmaking arises during discovery-time. This approach may be placing an overhead on the time required to complete the publication of a service advertisement, but it substantially reduces the time required to perform matchmaking at discovery-time, so it is considered particularly beneficial.

Three conditions must hold in order to claim that the new service advertisement can satisfy a service request: (i) the `InputDataSet` concept associated with the RFP must be subsumed by the `InputDataSet` of the AFP, (ii) the `OutputDataSet` of the RFP must subsume the `OutputDataSet` of the AFP, and (iii) the `TaxonomyEntity` concept associated with the RFP must subsume the `TaxonomyEntity` of the AFP. The interoperability-oriented rationale that these classification conditions reflect, and the way in which they collectively form a set of criteria for satisfactory matchmaking, is explained in section 4.

Finally, the Publication Manager maps the syntactic information extracted from the SAWSDL document and the semantic classification information derived by classifying the AFP onto appropriate UDDI data structures (`keyedReferences` to special-purpose `tModels`). Communication with the UDDI server module takes place through the UDDI Client API library, as illustrated in Figure 1. The mapping follows a well-defined methodology that is described in [13] and is beyond the scope of this paper to analyse. When the publication algorithm completes, a new semantic service advertisement has been created, registered with the UDDI registry, and is readily available for discovery.

3.2 Functionality of the Discovery Manager Module

The Discovery Manager Module provides interfaces for semantic matchmaking of a given service request against the published service advertisements, and for retrieving analytical information about records of advertisements and their providers.

---

[7] http://knoesis.wright.edu/opensource/sawsdl4j/
[8] http://owlapi.sourceforge.net/
[9] http://pellet.owldl.com/

The discovery query that initiates the semantic matchmaking process comprises two elements:

i. a URI pointing to some Request Functional Profile (RFP) that represents the characteristics of the Web service sought
ii. an optional system ID indicating the preferred service provider, i.e. the business information system that the service should originate from

The first step in the discovery procedure is to resolve the location of the RFP that is referenced by the provided URI. The RFP may be defined either within the FUSION Ontology that is shared by service providers and service requestors alike (i.e. be a shared RFP), or in some third-party ontology that imports and extends the FUSION Ontology (i.e. be a custom-built RFP). Depending on which of the two cases holds, the algorithm would follow a different discovery path:

i. If the RFP is defined within the FUSION Ontology, a syntactic, UDDI-compliant discovery query is generated and submitted directly to the UDDI server through the UDDI Client API library depicted in Figure 1.
ii. If the RFP is defined in a third-party ontology that is not shared with the service provider the Discovery Manager will load the ontology in which the RFP is defined to the DL Reasoner and compute the subsumption hierarchy.

Due to the shared ontology assumption that is valid in FUSION, the first case is the most typical type of discovery querying envisaged for the FUSION Semantic Registry, and is also the simplest and fastest type of matchmaking possible. Since the time-consuming process of concept classification has been already performed at publication-time, the computational complexity of discovery-time matchmaking for RFPs defined in a shared ontology is essentially as low as that of a conventional UDDI server.

The result of the discovery process is a list of advertisements complying with the matchmaking criteria captured by the RFP. If the optional system ID has been specified as part of the discovery query to indicate the preferred service provider, the registry uses it to filter-out services that are offered by systems other than the one specified. The ID is defined as an optional parameter in the discovery query, as it sometimes preferable to search for services that are offered anywhere within a service ecosystem, regardless of which business application exposes them.

4 Matchmaking Capabilities of the FUSION Semantic Registry

Due to the employed approach of OWL-DL-based service capability profiling and matchmaking, the FUSION Semantic Registry supports the evaluation of relevance among a service advertisement and a request at three distinct levels: (i) categorisation-level matching, (ii) message-level matching, and (iii) schema-level matching.

4.1 Categorisation-Level Matching

The end goal in this type of matchmaking is to determine if the categorisation value attributed to a service request is equivalent, more specific, or more generic than the one specified in some service advertisement. As an example consider the case of a Request Functional Profile (RFP) classified under *Supply Chain Management* services, and some Advertisement Functional Profile (AFP) classified under *Freight Costing* services, a subcategory of *Transportation* services that is classified under *Supply Chain Management* services. As already discussed in the previous section, the `TaxonomyEntity` concept associated with an RFP must subsume the `TaxonomyEntity` of the AFP in order to have a match. In this example this obviously holds since the category of *Supply Chain Management* services with which the RFP is associated is more generic (subsumes) the *Freight Costing* services category of the AFP.

4.2 Message-Level Matching

The end goal in this type of matchmaking is to determine the degree to which a service can produce the set of output data that the requestor wants to obtain, and the degree to which the requestor can provide the set of input data that a service needs to receive when invoked. Positive matchmaking in this respect is essential for guaranteeing flawless communication and interoperability among a chain of composed services. By referring to sets of input and output data, instead of request and response messages, we intend to abstract from the differences among *complex* and *atomic* Web services. In the case of atomic, non-transactional Web service operations, the set of input data trivially corresponds to an operation's request message, while the set of output data corresponds to its response message. In the case of complex, transactional services involving the invocation of numerous Web service operations to fulfil one goal, the set of input data corresponds to the superset of all sets of input data exchanged as part of request messages for the operations involved, while the equivalent holds for output data.

As a result, the degree of match among the inputs or outputs of an AFP and an RFP would be determined by the degree to which their respective `InputDataSet` or `OutputDataSet` contain common elements. To provide a formal definition of degree of match we adopt the set-theoretic model from the work of [14] and [15]:

i. Exact Match: The advertisement consumes (for input-matching) or produces (for output-matching) the data that is exactly specified in the request
ii. Plugin Match: The advertisement consumes or produces all data specified in the request, but also consumes or produces some irrelevant data
iii. Subsumption Match: The advertisement consumes or produces only some of the data specified in the request, and no irrelevant data
iv. Intersection Match: The advertisement consumes or produces only some of the data specified in the request, but also consumes or produces irrelevant data
v. Non Match: The advertisement consumes or produces none of the data specified in the request

When checking for input message compatibility the cases of exact and subsumption match are the only ones that can be considered safe for interoperability and thus satisfactory for positive matchmaking. In the rest of the cases the advertised service is not guaranteed to receive all the input data it requires, and thus run-time errors could arise. Similarly, if we were checking for output message compatibility, the positive matchmaking cases are exact or plugin match. As a negative match example, consider the case of an RFP representing a request for a shipment cost calculation service, having an `InputDataSet` that contains `Product` and `Customer` and an `OutputDataSet` that contains `ShipmentDetails`. Consider also an AFP with an `InputDataSet` that contains `Address`, `Product`, and `Customer`, and an `OutputDataSet` containing `ShipmentDetails` and `DigitalSignature`. Despite the fact that the `OutputDataSet` of the RFP subsumes the `OutputDataSet` of the AFP (i.e. the advertised service can offer more than what is being asked for), the `InputDataSet` concept of the RFP is not subsumed by the `InputDataSet` of the AFP (i.e. the advertised service asks for more than what can be provided).

4.3 Schema-Level Matching

The end goal in this type of matchmaking is to determine the degree to which the schema of some data parameter produced or consumed by an advertised service contains all the attributes specified in the corresponding schema of the request. When working under the assumption of a shared base ontology that can be specialised and customised for niche application domains through subclassing and applying quantification restrictions on properties, as in the case of the FUSION Ontology, the case may arise where different partners have chosen to extend a base ontology concept in different ways, thus creating potential interoperability problems. Figure 2 illustrates an example case in which the base concept of `Address` (depicted in the middle column) has been subclassed and specialised in two different ways, for modelling two different business applications.

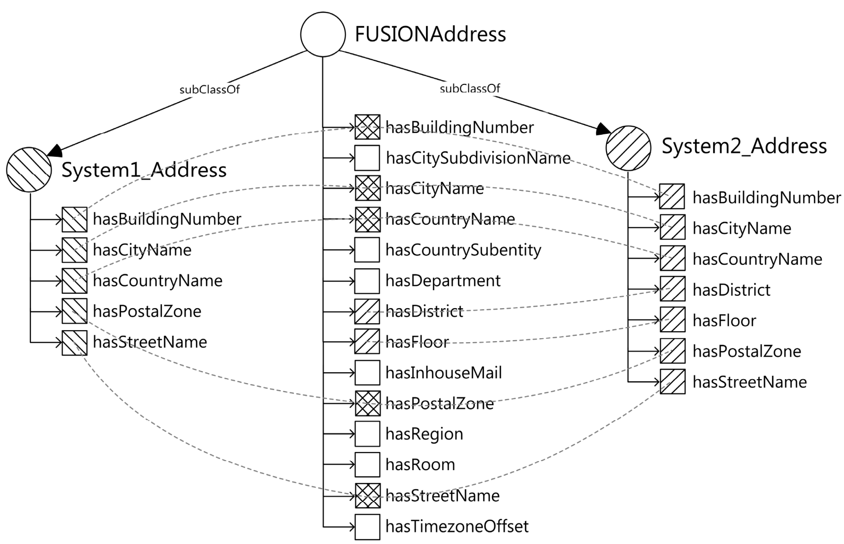


Fig. 2. Schema-level mismatch due to concept subclassing (excerpted from [5])

Although `System1_Address` and `System2_Address` are subclasses of the same concept, if they are used in the context of input and output data exchanged by the two systems, interoperability cannot be always guaranteed. The schema of `System2_Address` is more specific than that of `System1_Address`, since the first specifies more attributes than the schema of the latter. In fact, if `System1` was to consume a service exposed by `System2`, and the service requested to be provided with address information as input, `System2` could consume all of the data included within `System1_Address`, but still require some additional data (`hasDistrict`, `hasFloor`) that would not be provided, thus leading to potential problems during process execution. As in message-level matching, the cases of exact or subsumption match could be considered satisfactory for positive matchmaking when checking for input compatibility, while the rest of the cases could not. When checking for output compatibility the cases considered satisfactory for positive matchmaking would be exact or plugin match.

5 Related Work

Recent years have seen an increasing interest on the use of semantics to represent service capabilities and on the introduction of semantic matchmaking functionality to UDDI registries, and numerous works could be considered relevant to ours. In this section we however discuss only related works that build on the established Semantic Web Service frameworks of OWL-S [6], WSMO [7], and WSDL-S [8].

In [2] and [16] the authors propose that discovery in UDDI registries should be realised through semantic matchmaking of service capability descriptions that are expressed as OWL-S Profiles and mapped onto UDDI structures. They propose the incorporation of a matchmaking engine inside the UDDI registry, thereby necessitating the modification of the UDDI server's interface and implementation. In a subsequent work [17] a revised mapping among OWL-S Profiles and the UDDI data model is proposed, and an improved version of the matchmaking algorithm from [16] is presented. Semantic classification and indexing are performed at publication-time rather than discovery-time and as pointed out by the authors in the paper the proposed solution is incomplete since discovery-time classification is not allowed.

In [18] the authors build on the approach proposed in [16] and present a method that improves the effectiveness of service discovery in UDDI based on a two-stage service discovery process, combining syntactic and semantic search. The expressiveness of the semantics that the proposed matchmaking algorithm employs are in the range of RDFS and OWL-Lite, and as a result the proposed solution cannot be used for matchmaking over highly expressive schema descriptions (e.g. with arbitrary cardinality restrictions on properties). Similarly to the approach by Paolucci et al, the solution proposed in [18] also necessitates some changes to the API and implementation of the UDDI server.

In [3], and later in [19], the authors present an approach for publishing semantically annotated WSDL descriptions based on a methodology for WSDL-S to UDDI mapping. Annotations are stored in UDDI and discovery is performed based on a semantic request template that captures abstract service characteristics. To perform matchmaking the described platform implements a semantic reasoner based on the Jena API. The reasoner supports semantic entailments for OWL-Lite but does not fully support OWL-DL, and therefore the proposed solution has some limitations as the one in [18].

A number of discovery engine implementations have been also developed in the context of the WSMX Working Group [20] for supporting the three different discovery approaches that are put forward in WSMO [15]: keyword-based discovery, lightweight semantic discovery (based on WSML-Rule and WSML-DL), and heavyweight semantic discovery (based on WSML-Flight). The specific works however do not offer themselves to direct comparison with our work or the other approaches discussed in this section, since they do not attempt to provide semantic enhancements to UDDI but rather stand as independent WSMX environment components and are not integrated with UDDI.

6 Conclusions

To promote interoperability among service-oriented business applications and efficient business process integration, the FUSION project promotes the introduction of semantics to Web service discovery in UDDI registries. In this paper we provided an overview of how UDDI, OWL-DL semantics, SAWSDL annotations and DL reasoning are employed within FUSION to enhance service discovery, we presented the FUSION Semantic Registry architecture and provided a walkthrough of the main activities performed during service publication and service discovery. Moreover, we analysed the matchmaking capabilities of the FUSION Semantic Registry and discussed its applicability in practical terms for evaluating the degree of match among service advertisements and service requests at three distinct levels: categorisation-level matching, message-level matching, and schema-level matching. To the best of our knowledge this the first attempt to combine SAWSDL-based service descriptions with OWL-DL based service capability profiling and automated matchmaking through DL reasoning in a semantically extended UDDI registry.

Acknowledgments. Research project FUSION (Business process fusion based on semantically-enabled service-oriented business applications) is funded by the European Commission's 6th Framework Programme for Research and Technology Development under contract FP6-IST-2004-170835 (http://www.fusion-strep.eu/).

References

1. Bellwood, T., Bryan, D., Draluk, V., Ehnebuske, D., Glover, T., Hately, A., Husband, Y.L., Karp, A., Kibakura, K., Kurt, C., Lancelle, J., Lee, S., MacRoibeaird, S., Manes, A.T., McKee, B., Munter, J., Nordan, T., Reeves, C., Rogers, D., Tomlinson, C., Tosun, C., von Riegen, C., Yendluri, P.: UDDI Version 2.04 API Specification, UDDI Committee Specification (July 2002)
2. Paolucci, M., Kawamura, T., Payne, T.R., Sycara, K.: Semantic Matching of Web Service Capabilities. In: Horrocks, I., Hendler, J. (eds.) ISWC 2002. LNCS, vol. 2342, Springer, Heidelberg (2002)
3. Sivashanmugam, K., Verma, K., Sheth, A., Miller, J.: Adding Semantics to Web Services Standards. In: Proceedings of the 2003 International Conference on Web Services (ICWS 2003), Las Vegas, USA (June 2003)

4. Colgrave, J., Akkiraju, R., Goodwin, R.: External Matching in UDDI. In: Proceedings of the 2004 IEEE International Conference on Web Services (ICWS 2004), USA (July 2004)
5. Kourtesis, D., Paraskakis, I., Friesen, A., Gouvas, P., Bouras, A.: Web Service Discovery in a Semantically Extended UDDI Registry: the Case of FUSION. In: Camarinha-Matos, L., Afsarmanesh, H., Novais, P., Analide, C. (eds.) IFIP International Federation for Information Processing, Establishing the Foundation of Collaborative Networks, vol. 243, pp. 547–554. Springer, Boston (2007)
6. Martin, D., Burstein, M., Hobbs, J., Lassila, O., McDermott, D., McIlraith, S., Narayanan, S., Paolucci, M., Parsia, B., Payne, T., Sirin, E., Srinivasan, N., Sycara, K.: OWL Web Ontology Language for Services (OWL-S). W3C Member Submission (November 22, 2004)
7. Bruijn, J.d., Bussler C., Domingue J., Fensel D., Hepp M., Keller U., Kifer M., Konig-Ries B., Kopecky J., Lara R., Lausen H., Oren E., Polleres A., Roman D., Scicluna J., Stollberg, M.: Web Service Modeling Ontology (WSMO). W3C Member Submission (June 3, 2005)
8. Akkiraju, R., Farrell, J., Miller, J., Nagarajan, M., Schmidt, M.T., Sheth, A., Verma, K.: Web Service Semantics (WSDL-S). W3C Member Submission (November 2005)
9. Farrell, J., Lausen, H. (eds.): Semantic Annotations for WSDL and XML Schema (SAWSDL). W3C Recommendation (August 2007)
10. Bouras, A., Gouvas, P., Mentzas, G.: ENIO: An Enterprise Application Integration Ontology. In: 1st International Workshop on Semantic Web Architectures For Enterprises, 18th International Conference on Database and Expert Systems Applications, Regensburg, Germany, September 3-7 (2007)
11. McGuinness, D.L., van Harmelen, F.: OWL Web Ontology Language Overview, W3C Recommendation (February 2004)
12. Pokraev, S., Koolwaaij, J., Wibbels, W.: Extending UDDI with Context Aware Features based on Semantic Service Descriptions. In: Proceedings of the 2003 International Conference on Web Services (ICWS 2003), Las Vegas, USA (June 2003)
13. FUSION project Deliverable D3.1 – Specifications of the Integration Mechanism (April 2007), http://www.fusion-strep.eu/
14. Li, L., Horrocks, I.: A Software Framework for Matchmaking Based on Semantic Web Technology. In: Proceedings of the 12th International World Wide Web Conference (WWW 2003), Budapest, Hungary (May 2003)
15. Keller, U., Lara, R., Polleres, A., Toma, I., Kifer, M., Fensel, D.: WSMO D5.1 – WSMO Web Service Discovery (v0.1). WSML Working Draft (November 2004)
16. Paolucci, M., Kawamura, T., Payne, T.R., Sycara, K.: Importing the Semantic Web in UDDI. In: Proceedings of Web Services, E-Business and Semantic Web Workshop, Toronto, Canada, May 2002, pp. 225–236 (2002)
17. Srinivasan, N., Paolucci, M., Sycara, K.: Adding OWL-S to UDDI, Implementation and Throughput. In: Cardoso, J., Sheth, A.P. (eds.) SWSWPC 2004. LNCS, vol. 3387, Springer, Heidelberg (2005)
18. Akkiraju, R., Goodwin, R., Doshi, P., Roeder, S.: A method for semantically enhancing the service discovery capabilities of UDDI. In: Proceedings of the Workshop on Information Integration on the Web (IIWeb 2003), Acapulco, Mexico (August 2003)
19. Li, K., Verma, K., Mulye, R., Rabbani, R., Miller, J., Sheth, A.: Designing Semantic Web Processes: The WSDL-S Approach. In: Cardoso, J., Sheth, A. (eds.) Semantic Web Services, Processes and Applications, pp. 163–198. Springer, Heidelberg (2006)
20. WSMX (Web Service Modelling eXecution environment), http://www.wsmx.org/

Towards the Aggregation of Security Requirements in Cross-Organisational Service Compositions

Michael Menzel[1], Christian Wolter[2], and Christoph Meinel[1]

[1] Hasso-Plattner-Institute, University of Potsdam, Germany
{michael.menzel,meinel}@hpi.uni-potsdam.de
[2] SAP Research, CEC Karlsruhe, Germany
christian.wolter@sap.com

Abstract. The seamless composition of independent services is one of the success factors of Service-oriented Architectures (SOA). Services are orchestrated to service compositions across organisational boundaries to enable a faster reaction to changing business needs. Each orchestrated service might demand the provision of specific user information and requires particular security mechanisms. To enable a dynamic selection of services provided by foreign organisations, a central management of static security policies is not appropriate. Instead, each service should express its own security requirements as policies that stipulate explicitly the requirements of the composition. In this paper we address the problem of aggregating security requirements from orchestrated services. Such an aggregation is not just the combination of all security requirements, since dependencies and conflicts between these requirements might exist. We provide a classification of these dependencies and introduce a conceptional security model enabling a classification of security requirements to reveal conflicts. Finally, we propose an approach to determine an aggregation of security requirements in cross organisational service compositions.

Keywords: SOA, Service Composition, Security, Security Policy.

1 Introduction

Service-oriented Architectures (SOA) facilitate the interoperable and seamless interaction of service consumer and service provider to meet the consumer's needs by the service's capabilities. The standard set of Web Service technologies, such as WSDL, UDDI, and SOAP provides the means to describe, locate, and invoke a Web Service based on XML. The independent nature of the services, with respect to operating systems and system architectures, facilitate a composition of different services. In fact, service composition is one of the success factors of Service-oriented Architectures to enable the flexible integration of services provided by independent business partners.

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 297–308, 2008.

However, the seamless and straightforward integration of cross-organisational services conflicts with the need to secure and control the access to provided services. Each service may support different security mechanisms and may require different pieces of user information for access control. Since the user of the composed service might be unknown to the provider of the orchestrated services, the establishment of an identity federation is a basic necessity to provide required information across domain borders. In an identity federation all parties are willing to rely on assertions representing claims about users. These claims are issued by a trusted identity provider that manages a digital identity of the service user. For instance, a credit card company can assert the user's name and his credit card information in an encrypted security token (e.g. SAML [12]). A user can request this token on demand of a service's requirements and include this token in a SOAP message to invoke a service. Windows CardSpace [3] is one example for a client technology to manage digital identities from various identity providers. Based on the requirements of a service, CardSpace acts as an identity selector and enables a user to choose from a set of identity providers that can assert the required set of claims. Since this technology is founded on the WS-* protocols, WS-Policy [5] is used to express the requirements.

Security policies expressing requirements for the users of a service are typically generated when the service is deployed, and are assigned to the service statically. This approach is useful in a single domain leveraging a central policy management, but it is not feasible in the context of service compositions containing services from independent organisations. The security requirements of the composed service do not only depend on local security requirements, but also on security policies and service level agreements specified by the orchestrated services. In consequence, the security policy of the composed service might have to be adapted when a service from another service provider is mapped to the composition. This is especially an issue in service compositions that dynamically select services based on non-functional properties. Moreover, it must be considered that there might be dependencies between different security requirements. Requirements might interact, contradict or conflict, so that a simple combination of all requirements is not sufficient.

Current research approaches [2,1] are based on a semantic matching of service's security preconditions to create a service composition that can be executed with the user's security capabilities. These approaches are focused on enabling an automatic composition under the restriction that the user must state in advance which security credentials and mechanisms he is able to support. This solution disregards privacy concerns, since it conflicts with the conception that the user should control the usage of his identity information by selecting an appropriate digital identity. Moreover, former approaches did not consider dependencies between security requirements that may result in an insecure policy of the service provider.

In this paper we propose a method to aggregate and verify security requirements for cross-organisational federated service compositions. Therefore, we provide:

- a classification of dependencies and effects between requirements (*requirement interactions*) that must be resolved when aggregating a consistent set of security requirements.
- a conceptual security model that describes entities (e.g. security goals, policies and security mechanisms) and their relationships regarding security requirements independently from technical aspects.
- an approach to determine a consistent aggregation of security requirements for a service composition

The rest of this paper is organised as follows. In Section 2 we present a travel agency scenario to illustrate dependencies of requirements in a service composition and a definition of interaction classes between security requirements. In Section 3 we introduce our conceptional security model to distinguish different types of security requirements. Finally, we describe in Section 4 how our model can be used to determine the interaction classes of security requirements and how security requirements can be aggregated in a service composition. Section 5 provides an overview about related work, while the final section concludes this paper.

2 Interactions of Security Requirements

In this section we present an example to clarify the aggregation of security requirements in a service composition. Moreover, we provide a complete list of possible interactions between these requirements that have to be considered to compute a secure aggregation.

2.1 Travel Agency Service Scenario

Consider the example of a composed travel agency service that is capable to perform a hotel and a flight reservation as shown in Figure 1. We assume that the request to this service has to include all needed functional parameters, e.g. preferred hotel, room types, arrival and departure time, and preferred type of flight. Based on this information, the service is able to book the preferred combination of hotel and flights. The service composition contains four independent services whereas the services a_2, a_3, and a_4 are provided by business partners of the service a_1. The first service a_1 validates and verifies the functional service parameters and invokes the service a_2, which performs the reservation of the desired hotel. Finally, depending on the preferred type of flight, either the service a_3 is called to book a budget flight or the service a_4 is called to book a regular flight.

In addition to these functional parameters, identity related information must be provided to complete the booking process successfully. The services a_2, a_3, and a_4 require the name and address of the user as well as his credit card information for payment. To secure the exchanged information, different security mechanisms are needed to ensure confidentiality and integrity. While the services a_3 and a_4 require the user to decrypt the information using the AES algorithm,

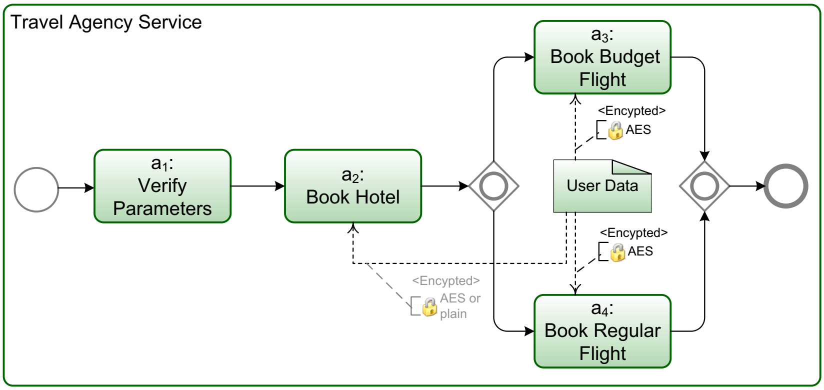


Fig. 1. Travel Agency Service Composition

it is optional for service a_2. WS-Policy can be used to express the requirements of the individual services.

Although the policies of the basic services are well defined, it is challenging to generate a secure policy for the service composition. The collection and simple combination of all security assertions is not sufficient, since independent requirements might interact. In our example, the simple combination of the basic policies would allow that two security tokens conveying the same pieces of identity information are added to the service request with different requirements for encryption. One security token with the users name and address would be encrypted with AES, while the other token with the same information can be added to the same message without encryption. It is obvious that this is a security breach, but what if service a_2 supports a weaker encryption mechanism (e.g. DES) or AES with a shorter key length? The security level regarding the identity information would depend on the weakest security mechanism. Another possibility is that different services require different mechanisms to guarantee integrity. How can be ensured that they provide the same level of security?

2.2 Classification of Requirement Interactions

A secure service complies with a set of security goals $G_i = \{g_1, \ldots, g_m\}$ (e.g. authorisation or confidentiality) requiring a certain behaviour or information from a client that wants to access the service. These expectations are expressed as a set of requirements $R_i = \{r_{i1}, \ldots, r_{im_i}\}$ associated with this service. To guarantee a consistent and valid set of requirements, interactions (e.g. conflicts) between these requirements must be considered. A *requirement interaction* is the effect that two requirements have on each other. We introduce a classification of security requirement interactions in this section based on former work on QoS interactions described by Wohlstadter et el. [13]. A requirement has a positive or negative impact that increases or decreases the service's security. This impact is expressed by the function Ψ. Based on the comparison of the individual security impact of two requirements r_1 and r_2 with the combined one, eight classes of interaction can be revealed:

- *Independent* - $\Psi(\{r_1, r_2\}) = \Psi(\{r_1\}) + \Psi(\{r_2\})$
  The requirements are totally independent and do not interfere with each other. The combined security impact is the sum of the individual ones.
- *Equivalent* - $\Psi(\{r_1, r_2\}) = \Psi(\{r_1\}) = \Psi(\{r_2\})$
  Two requirements will be equivalent, if they have the same effect providing the same level of security. It is optional to use both requirements together, although it might be advantageous to gain a greater flexibility by offering equivalent alternatives to the service consumer.
- *Prevent* - $\Psi(\{r_1, r_2\}) = \Psi(\{r_1\})$
  The requirement r_1 will prevent r_2, if r_2 has no impact on the security. Consider the aforementioned travel agency scenario describing a service invocation that contain the same pieces of information in an encrypted and a plain style. The unsecure requirement prevents the high security requirement to effect the communication.
- *Restrict* - $\Psi(\{r_1, r_2\}) < \Psi(\{r_1\}) + \Psi(\{r_2\})$
  Both requirements lower the impact of the other one. The combined impact is usually greater than the individual ones, but less than the sum of those.
- *Complements* - $\Psi(\{r_1, r_2\}) > \Psi(\{r_1\}) + \Psi(\{r_2\})$
  The requirements will be complementary, if the combined security impact is larger than the individual security level. It is more secure to use both requirements together.
- *Require* - $\Psi(\{r_1, r_2\}) > 0, \Psi(\{r_1\}) \leq 0)$
  A requirement r_1 will require r_2, if the combined security impact is positive, while the individual impact of r_1 is negative. Both requirements must be used together.
- *Conflict* - $\Psi(\{r_1, r_2\}) < 0, \Psi(\{r_1\}) \geq 0 \wedge \Psi(\{r_2\}) \geq 0$
  Two requirements will conflict, if their combined impact is negative, although their individual impacts are positive. These requirements must not be deployed together.
- *Exclude* - $\Psi(\{r_1, r_2\}) = \Psi(\{r_1\}) \vee \Psi(\{r_1, r_2\}) = \Psi(\{r_2\})$
  The requirements r_1 and r_2 represent an excluding alternative that may be selected in dependency on functional or non-functional parameters of a service.

Although the possibility of interacting security requirements for a single service has to be considered, this problem is much more serious in terms of service compositions. In a service composition there are multiple sets of security requirements $R_i = \{r_{i1}, \ldots, r_{im_i}\}$, $m_i \in \mathbb{N}$ involved that are associated with each service i. Each service can have its own understanding of security and states its own requirements regarding the same security goals. Therefore, if R_{comp} is the set of consistent security requirements for the service composition satisfying a set of predefined security goals, then $R_{comp} \subseteq \bigcup_{i=1 \ldots n} R_i$. Due to possible interactions between security requirements, it is likely that R_{comp} is unequal to the union of all requirements.

3 Security Model

Security requirements are expressed by security policies, usually in a very technical and policy language dependent way. To determine the interaction type of two requirements, we need a security model that abstracts from technical details. The model must reveal all security aspects in an SOA landscape and the relationship among affected entities. Therefore, our conceptual security model describes basic security goals and outlines the relationship to specific security attributes and mechanisms.

3.1 Specifying Security Goals

The abstract concept of security can be defined precisely by specifying a set of security goals [11]. Although these goals can be further specialised, subdivided or combined, we will focus solely on basic goals in this paper:

1. *Confidentiality* provides protection against the unauthorised notice of stored, processed, or transferred information.
2. *Integrity* ensures the properness (intactness, correctness, and completeness) of information (data integrity). Transferred, processed, or stored data must not be modified with proper rights and - in economic terms - modifications must correspond to business values and expectations.
3. *Authentication* ensures the credibility of information - such as a claimed identity - by confirming this information as authentic.
4. *Authorisation* is the process of granting rights to participants to perform an interaction, for instance to access a resource.
5. *Traceability and Auditing* provide verifiability regarding all performed actions in an information processing system. This can be related to simple logging mechanisms, but also to monitoring as real-time auditing e.g. in intrusion detection systems.
6. *Availability* ensures that data, resources and services, which are needed for the proper functioning of a system, are available at each point in time regarding the requested quality of service.

These goals can be related to various entities in a Service-oriented Architecture. The relations among security goals and affected entities are typically described by *Constraints* that are composed in a security *Policy* as indicated by Figure 2. A subset of constraints can be exposed as requirements to service user. Since some constraints regulate the internal functioning of the service, they are not revealed to the service user.

The basic entity in such a model is an *Object*. We define an object as an entity that is capable of participating in an *Interaction* with other objects. This interaction will always lead to an *Effect*, which can comprise the provision of information or the change of state in a system. The effect can, but does not need to be, related to the object that initiated the interaction. For example, one object could be an application and another object could be a service to store data. The process of accessing this service would be the interaction resulting in the effect that data is stored or some information is returned to the application. This conception is close to the model that is described in the OASIS reference model for SOA [10] and enables a straight mapping to this model.

Each object is related to a set of attributes describing its meta information. For instance, if the object represents a user, attributes, such as name, email address, age, etc. will be assigned. Altogether, policy constraints always refer to a set of objects, a particular set of objects' attributes, and optionally a set of interactions and effects that are related to the objects. Based on these relations, specific constraints for particular security goals can be defined. These specific constraints define requirements for associations between the entities with regard to the particular security goals.

As shown in Figure 2, constraints specify security mechanisms that enforce or guarantee the defined constraint. For instance, a confidentiality policy usually specifies an algorithm (e.g. DES) that must be used to guarantee this requirement.

3.2 Enforcing Security Constraints

In our model a *Security Mechanism* is designed to characterise techniques that are used to enforce security constraints (cf. Figure 3). It provides the foundation to specify a comprehensive ontology for security mechanisms, see [4].

Besides security mechanisms, a *Credential* represents another important entity in our model that subsumes evidences used by security mechanisms. A detailed classification of security credentials was presented by Denker *et al.* [4]. In this work they introduced an ontology that divides credentials in simple credentials (e.g. key, login, certificate) and composed credentials (e.g. Smart Card, SAML, WS-Security Token) that contain a set of simple credentials.

Based on the given security policy model (cf. Figure 2), we defined semantics for specific types of *Constraints*, each guaranteeing one of the security goals listed above. Each constraint is related to a specific set of entities and define

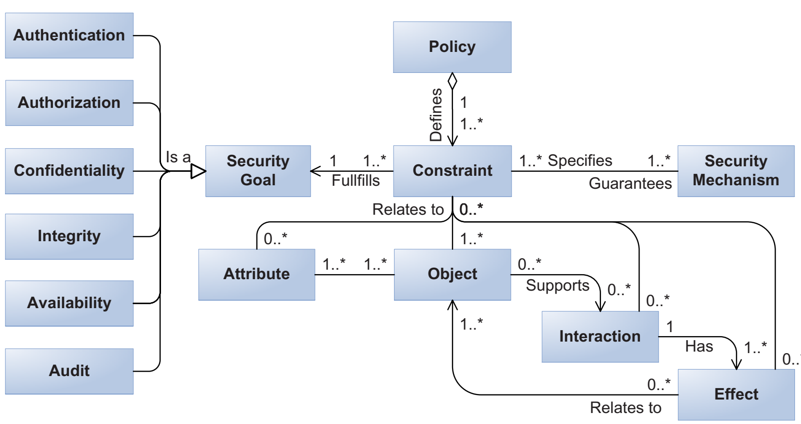


Fig. 2. Security Policy Model

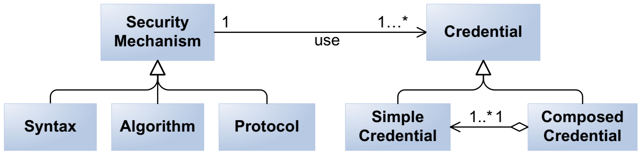


Fig. 3. Security Mechanisms Model

rules restricting particular associations between those entities. These rules must be enforced by security mechanisms and credentials. We will not elaborate upon the constraint models during the course of this paper, since these details have only a minor impact on the relationship among security requirements.

4 Aggregation of Security Requirements

In the previous section we introduced a security model that provides semantics for all security aspects related to a service. It has revealed that policy constraints (describing security requirements) are based on two basic ontologies: *mechanisms* to guarantee a certain security goals and *credentials* to provide required information. In this section we will present how this model can be used to determine the interaction class of security requirements. Finally, we will describe an approach to aggregate security requirements concerning a service composition.

4.1 Determining a Consistent Aggregation of Security Requirements

Determining the interaction class of multiple requirements demands an evaluation of these requirements regarding the different aspects in our model and their relationships. First of all, it must be considered, if two requirements belong to the same class of security goals. Second, it is important, whether they refer to the same set of entities in terms of objects, attributes, and object interactions.

A requirement interaction will belong to the class

- *Independent*, if the requirements refer to different entities.
- *Prevent, Restrict, or Equivalent*, if the requirements refer to the same security goals and relate to the same entities. It is necessary to compare the security mechanisms and security credentials specified in the requirements to determine the interaction type precisely. A comparison of security mechanisms can be realised similar to the concept of trust indicators, which was proposed by Haller *et al.*[6]. Trust indicators are based on specific metrics to measure and compare security properties concerning service level agreements.
- *Require, Complements, or Conflict*, if the requirements refer to different security goals and relate to the same entities. In general, these requirement interactions are inherited from dependencies between the associated security goals. The provision of authentication and authorisation information might require confidentiality and integrity to secure this information. Moreover,

there might be contracting security goals that lead to conflicting requirements (e.g. monitoring and confidentiality).
– *Exclude*, if there are conditions defined in policy constraints that limit the application to one requirement exclusively, regardless of the related security goals or referred entities.

Security constraints defined in existing standards (e.g. WS-Policy) can be mapped to our model to ensure the policy's consistency. Restricting or preventing requirements can be eliminated if they are optional considering the revealed dependencies. In general, conflicting requirements demand a conflict resolution strategy to decide which requirements should be preferred. For instance, confidentiality requirements could take preference over monitoring requirements.

4.2 Aggregating Security Requirements in a Service Composition

To enable an aggregation of security requirements regarding a service composition, we have to consider that

– each service has its own security requirements that might cause complex dependencies between the requirements of different services.
– the execution of services results in an effect that can satisfy the requirement of a successive service. This requirement does not have to be satisfied by the user.
– the service composition might include xor-splits and -joins depending on conditions based on functional parameters. This might result in different sets of requirements that exclude each other.

Results from research work [9] about the generation of semantics of service compositions provide a suitable foundation for the determination of the composition's security requirements. These approaches are based on the aggregation of semantic preconditions of basic services that correspond to our requirements. Meyer [9] introduces a formal workflow model based on petri nets and describes how to express the semantics of services inside this petri net. To calculate the functionality of the aggregation, Meyer introduced graph algorithms that work by recursively defining the state for the markings in the workflow net. Using our security model, and comprising ontologies for *security mechanisms* and *credentials*, we are able to translate policies into semantic preconditions for services. These preconditions can be used with the description of the service composition as input to the aforementioned algorithms to generate the requirements of the service composition expressed as Boolean formula. This output can be checked for consistency as described in the previous section and be mapped to a policy again.

5 Related Work

Our conceptional data model for service based systems is mainly driven by the semantic security annotation approach for web services proposed by Denker *et al.*

in [4]. These annotations are used to describe the security capabilities of web services. A reasoning engine is used to perform a security matchmaking between service providers and requesting agents. While providing suitable security ontologies for security mechanisms and credentials, they do not consider the relationship to policy constraints and possible interactions.

Only a few approaches have been published so far that intent to enable the dynamic composition of services regarding security constraints. Carminati *et al.* described an approach to compose web services based on security requirements of web service consumers and web service providers in [2,1]. Semantic matchmaking is used to create a service composition under security constraints. This approach assumes that service consumers are willing to specify their requirements and capabilities in advance. Dependencies between security requirements are not considered in this work.

Cheikh *et al.* proposed technique for automatic web service compositions in trust-aware communities based on reduction to satisfiability in propositional dynamic logic. Their work is focused on access control and authorisation constraints and do not consider other security goals.

The abstract description of security properties has also been addressed in previous work. However, there is no comprehensive approach that relates all security aspects in a SOA to security goals. Huang presents in [14] a framework for semantic descriptions for web service security constraints. His approach intents to enable a reasoning over non-functional properties and the integration of business rules. Although a general framework to perform a reasoning is described, the paper does not provide a concrete security ontology to describe security constraints and their relationship to security goals. Moreover, an aggregation of security constraints is not considered.

Jürjens presented in [7] the UMLSec extension for UML to express security relevant information within a system specification diagram. The focus of UMLSec lies primarily on the modelling of communication-based security goals, such as confidentiality, for software artefacts rather than a more general approach to modelling a variety of security goals and their relationship.

6 Conclusion

In this paper we presented an approach to aggregate security requirements of services that are composed across organisational boundaries. Each orchestrated service provides its own security policies that stipulate the security requirements of the composed service. Our vision is to facilitate the dynamic composition of services by enabling the calculation of a consistent requirement aggregation that can be communicated to the service user as a security policy. The provision of such an aggregated security policy enables the service user to securely select an appropriate digital identity representing the required information.

We showed that interactions may occur between security requirements exposed by services from different trust domains. We provided a definition and a complete classification of requirement interactions based on an abstract model.

To distinguish these requirements, we introduced a conceptual security model to add semantics to different types of security constraints. Our model revealed the structure of these constraints that describe requirements, their relationship to high level security goals and their dependency to security mechanisms and credentials as basic ontologies. Finally, we described how our model can be used to determine and resolve the interaction classes and introduced an approach to aggregate security requirements based on the computation of preconditions in semantic workflows.

In contrast to current research approaches that are focused on the dynamic composition of services with the intention to meet predefined capabilities of a service user, we introduced a solution that does not require the user to expose all security capabilities in advance. Moreover, the interaction of requirements is not considered by former semantic approaches.

6.1 Future Work

We stated that our security model is an approach to describe requirements and their interactions. There is an ongoing effort to map our security model to WS-Policy and it must be proved that our model can be used to determine the interaction classes of security requirements. Moreover, we presented a basic approach to resolve requirement conflicts in this paper. More complex strategies can be integrated as it has been presented in research work concerning the resolution of policy constraint conflicts [8]. Finally, we have to prove that the petri net models used to calculate the semantic precondition of workflows can be extended with our model to determine the aggregation of security requirements. These topics will be addressed by future work.

References

1. Carminati, B., Ferrari, E., Hung, P.C.K.: Web Service Composition: A Security Perspective. In: WIRI 2005, pp. 248–253 (2005)
2. Carminati, B., Ferrari, E., Hung, P.C.K.: Security Conscious Web Service Composition. In: ICWS 2006, pp. 489–496 (2006)
3. Chappel, D.: Understanding Windows CardSpace (April 2006)
4. Denker, G., Kagal, L., Finin, T.W., Paolucci, M., Sycara, K.P.: Security for DAML Web Services: Annotation and Matchmaking. In: ICWS 2003, pp. 335–350 (2003)
5. Della-Libera, G., Gudgin, M., et al.: Web Services Security Policy Language (WS-SecurityPolicy). Public Draft Specification (Juli 2005)
6. Haller, J., Wolter, C.: Trust Indicator Integration into SLAs for Virtual Organisations. In: eChallenges 2007 Conference (2007)
7. Jürjens, J.: UMLsec: Extending UML for Secure Systems Development. In: Jézéquel, J.-M., Hussmann, H., Cook, S. (eds.) UML 2002. LNCS, vol. 2460, pp. 412–425. Springer, Heidelberg (2002)
8. Jaeger, T., Sailer, R., Zhang, X.: Resolving constraint conflicts. In: SACMAT 2004, pp. 105–114. ACM, New York (2004)
9. Meyer, H.: On the Semantics of Service Compositions. In: Proceedings of The First International Conference on Web Reasoning and Rule Systems (RR 2007) (2007)

10. MacKenzie, M., Laskey, K., McCabe, F., Brown, P., Metz, R.: Reference Model for Service Oriented Architecture 1.0. In: OASIS Committee Specification (February, 2006)
11. Charles, P.: Pfleeger and Shari Lawrence Pfleeger. Security in Computing. Prentice Hall Professional Technical Reference (2002)
12. Ragouzis, N., Hughes, J., Philpott, R., Maler, E.: MalerSecurity Assertion Markup Language (SAML) V2.0 Technical Overview(2006)
13. Wohlstadter, E., Tai, S., Mikalsen, T., Rouvellou, I., Devanbu, P.: GlueQoS: Middleware to Sweeten Quality-of-Service Policy Interactions. In: ICSE 2004, pp. 189–199 (2004)
14. Huang, D.: Semantic Descriptions of Web Services Security Constraints. In: SOSE 2006, pp. 81–84. IEEE Computer Society, Los Alamitos (2006)

Industrialized Generation of Financial Services by Component-Based Information Systems

Benjamin Olschok and Günter Schmidt

Saarland University, Department of Information and Technology Management,
66041 Saarbrücken, Germany
{bjo,gs}@itm.uni-sb.de

Abstract. Industrialization is an innovative trend not only in information systems development but also in the financial services sector. In order to support modularized business services by component-based information systems, we apply models and methods from manufacturing like bill of materials and process plans. These component-based information systems are implemented based on a service-oriented architecture. For each service element a software component is defined. Mass customization and generation of product variants are achieved by composition of these components. As a proof of concept, Life Charts, a new advisory approach in personal financial planning, is implemented based on financial web services. The approach is evaluated using the criteria performance, development cost, maintenance cost, and capability to handle complex compositions.

Keywords: CBSE, SOA, Web Services, Industrialization, Mass Customization, Personal Financial Planning.

1 Introduction

The objective of industrialization in the financial services sector and in software development is to overcome the problems of standardized business services and inflexible information systems. Individualization of such business services and information systems supporting these services is essential. Cost pressure forces to keep down the individualization effort. Hence, models and methods of mass production like mass customization (MC) are applied to computer-supported financial services.

MC combines mass fabrication - characterized by high number of pieces and low unit costs - and individual customization - characterized by low number of pieces and high unit costs [1]. Production systems are leveraged to increase variability and flexibility of mass production and reduce unit costs and throughput time of individual customization of goods. The objective of MC of services (cf. [2], [3]) is to vary services according to individual requirements. Costs should be kept down at the level of standardized and mass fabricated services. The configurability of the services is obtained by prefabricated elements and rules describing how these elements can be combined (cf. [4]); these considerations are in analogy to configurability of goods. In

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 309–320, 2008.

contrast to the research in MC of goods, only few results for the MC of services exist; this holds especially for financial services (cf. [2]).

The European and especially the German financial services sectors are exposed to a transformation process caused by earnings crisis, intensive cost pressure and imprecise business models. Industrialization seems to be a trend in the financial services sector to cope with this change and future development (cf. [5], [6]). Despite simplification, standardization and automation of processes, process management and information systems are required to support industrialization in analogy to manufacturing. In combination with middleware technologies like service oriented architecture (SOA), quality and productivity are expected to increase [6]. From an industrialization perspective the current state in the financial services sector is that work flows are carried out manually supported by automated machines. So far, most efforts of improvement concentrate on processes within a single organizational unit showing a low division of labor. Information technology enables the division of labor in complex processes by modularization. Component-based software supports this by collaboration of components via networks.

A business service answers a specific question of a customer in an application domain. A (software) service can fulfill simple requests up to complex business processes [7]. A service provides the functionality of a software component via a network by defined interfaces (cf. [8], [9]). (Software) components are "prefabricated, pretested, self-contained, reusable software modules – bundles of data and procedures – that perform specific functions" [10].

Based on manufacturing, bill of materials and process plans are also necessary for the composition of business services supported by software components. In a case study of personal financial planning with Life Charts, we derive bill of materials and process plans by the component identification approach proposed in [11]. We demonstrate how demands of customer segment specific services can be met and product variants are generated by applying findings of manufacturing. Therefore, we propose to connect modular structured services directly with component-based software. MC by composition of business and corresponding software services is used to develop flexible information systems.

The paper is organized as follows. In Section 2 the research problem is formulated and related work is discussed. Section 3 illustrates how methods of manufacturing can be transferred to component-based information systems development. SOA is proposed to support composition of software component-coupled modularized services to business processes. Section 4 introduces the personal financial planning approach of Life Charts and presents the web service based implementation. Mass customization and generation of product variants of the component-based financial services are outlined. Section 5 concludes with a short summary and suggestions for future work.

2 Problem Formulation and Related Work

Information systems are used to answer questions arising in an application domain. In the area of PFP questions are related to the financial status of an individual. There exist standard questions which appear in different variants. Consider e.g. the standard question if it is feasible for an individual to purchase a house. Variants of this question might be (a) is it feasible for an individual to purchase a house from its savings and (b) which costs of a loan can be accepted by the individual when buying a house. (a) is the variant which considers the feasibility of buying a house from a savings perspective and (b) is the variant which considers the feasibility of buying a house from a loan perspective. We assume that standard questions to be answered and software components to answer them are given. The problem is to find answers for a given set of variants of questions related to the standard questions such that the number of required components not yet element in the component set is minimized. The solution has to be capable to model and implement complex compositions answering variants of the standard questions. Further requirements arise from the comparison to non-distributed software. The performance has not to be perceivably affected by the composition. The total costs of development and maintenance have not to exceed the costs of a non-distributed approach.

Assembly of software components or (web) services is widely discussed in literature (cf. e.g. [7], [8]). In [12] a component-based approach is applied to questions in the financial services industry related to trading of assets by utilizing a given service set. The question to be answered is related to the effectiveness of a trading strategy. In a case study the simulation of a particular trading strategy is realized by composition of an exchange service and a trading data service. The exchange service matches buyer and seller orders. The trading data service provides historical and real-time information about quotes etc. Services are denoted in the Business Process Management Notation (BPMN). Single activities of the services are connected via interfaces. Functionality of legacy systems is encapsulated and provided as a web service. The case study shows feasibility and is evaluated addressing quality attributes like performance, development and maintenance effort and the capability of expressing complex business processes. Minimization of the number of services is not considered. The composition of the two mentioned web services is described by the use of a modelling tool. However, modeling realistic situations leads to complex and partly confusing diagrams. According to [12] the complexity refers to the limitations of BPMN. Due to the indirect access via the interfaces of the activities of the web services, the generation of product variants is hindered. Some performance loss of the SOA implementation is observed. E.g. the transactions rates of the exchange service are measured. The access via a web service causes around 10% maximum performance degradation compared to direct invocation of the legacy functionality. Development effort concentrates on the encapsulation of the functionality as web services and their composition. Development costs of new functionalities are not significantly increased by the encapsulation. Moreover the implementation of direct invocations is more complex than the implementation of the invocations in the web service composition. The latter invocations have not to be frequently changed, which results in a reduced maintenance effort.

3 Industrialization of Information Systems Development

Findings of engineering science have been in the focus of literature about component-based software engineering (CBSE) since the late sixties (cf. [13], [14]).

Industrialization of software engineering is a lately discussed topic (cf. [15]). We define industrialization by transferring well proven models and methods of manufacturing (cf. [16]) to software and information systems development. Industrialization requires modularization of the application domain into components. These components can afterwards be assembled to products answering specific questions of an application domain. The products correspond to goods in manufacturing, assembled of components and component parts.

The core of industrialization of services is the *composition model*. It consists of the bill of materials and the process plan. The bill of materials determines the components required in order to fulfill a particular business service; the process plan represents the collaboration of components. Following the approach proposed in [11], components can be identified by the use of functional abstraction. In the area of PFP a service is an answer to a specific question related to the financial status of an individual. The process plan determines the assembly schedule which provides information about the sequence of the components. Detailed steps to fabricate a service do not have to be specified because services are invoked via the interfaces of the already existing components. In manufacturing MC aims to increase flexibility in generation of modularized product variants. Customer specific questions of the application domain can be answered by a modification of the composition models. Adding and removing components or changing the assembly sequence enable the generation of product variants in order to answer variants of standard questions.

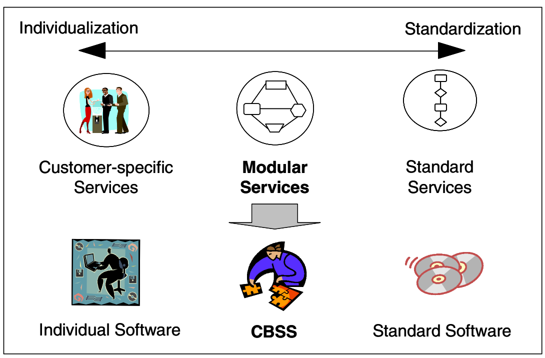


Fig. 1. Component-based Software and Modular Services

Component-based software qualifies to support MC of services due to direct coupling of the service elements to the components. Such software can be configured according to lists of variants and assembly schedules of variants. We define the direct connection of services to component-based software as CBSS (Component-based Software for Services). Note, that the resulting configurability exceeds the MC capabilities of reference models (e.g. [17]). The parts itself (by configurable software components) as well as the assembly schedule (by composition) can be diversified. The concept of modular structured services (cf. [3]) is extended by software aspects. The CBSS approach is positioned between expensive individual software and inflexible standard software (cf. Fig. 1).

Service-oriented Computing (SOC) is based on services as essential elements for information systems development. SOC relies to SOA realizing platform-independent composition of distributed applications by interacting services, standard interfaces and protocols. SOC and CBSE promise almost the same benefits: reduction of development time and costs as well as flexibilization of the information system (cf. e.g. [7], [8]). Principles like interoperability, loose coupling and reusability overlap with the demands on MC-suitable services as mentioned in [3]. The deficiencies of composed information systems are increased development effort compared to monolithic solutions and the effort to administrate the components or services. Accesses and invocations are realized via a network and interfaces. Hence, the amount of communication is increased, which results in some performance loss. In total, the benefits of CBSE/SOA like flexible processes, simplified implementation of business logic and reduced maintenance effort due modularization dominate.

Web services implements SOA by de facto technology standards. Web services are currently the most popular and promising realization of SOC [18]. The Business Process Execution Language for Web Services (BPEL) implements the composition of web services. The composition possibilities include linear sequences, decisions, loops and error-handling [19]. From the economical perspective, interoperability and flexibility increase due to the standardization [9].

4 Personal Financial Planning with Life Charts

Personal Financial Planning (PFP) is a cost- and time-intensive financial advisory service. PFP describes a continuous advisory process, gathering financial data and answering questions about the financial status in interaction with the client [20]. The objective of PFP is to identify improvements, to supervise financial actions and to monitor performance of action planning. Liquidity planning, risk management - including retirement planning – and asset allocation taking legal and tax aspects into account are the core tasks of PFP.

Life Charts [20] is a tool for planning, controlling and monitoring personal finance. This new advisory approach enables integrated consideration of as-is analysis, to-be concept and action planning. The time dependent financial aspects of a human life are represented by functions. These functions correspond to the required (demand) and the available (supply) funds of life (cf. Fig. 2).

The *required savings* $R(t)$ quote which amount of savings is required at the current age t up to the end of life at age T in order to afford the accumulated amount of nominal life expenditures in advance. When considering positive interest rates the amount of required savings is always smaller than the required capital. Under the assumption of the after-tax return rate r the calculation of $R(t)$ is according to (4.1).

$$R(t) = \sum_{i=t_0+1}^{T} w_i /(1+r)^{(i-t)} \tag{4.1}$$

The function of the *existing savings* $S(t)$ is known from the actual financial status. The existing wealth represents the nominal available amount of assets at time t. Returns on investment and amounts saved are achieved at the end of a period. Amounts saved (*b*) increase per period by the factor *h*.

$$S(t) = S_0 * (1+r)^{t-t_0} + b * \sum_{i=t_0+1}^{t} (1+h)^{(t-i+1)} * (1+r)^{(t-i)} \tag{4.2}$$

The intersection point of the required savings *R*(*t*) and the existing savings *S*(*t*) is the point of Financial Freedom (*FF-point*). Acting on the assumptions of the amounts saved, the inflation, tax and return rate, at this point of time the existing savings cover all future life expenditures ($S(t) \geq R(t)$). Reaching the FF-point means all future life expenditures can be financed by the existing savings.

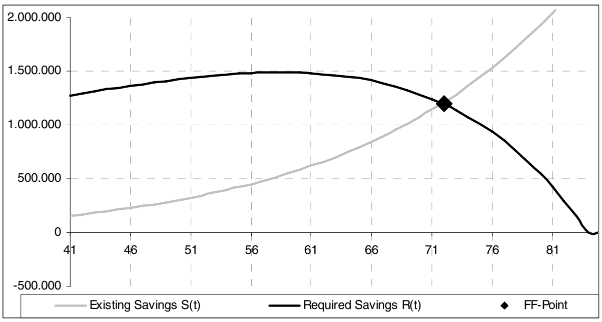


Fig. 2. Life Charts and FF-Point

Software supporting the advisory process of PFP is a core factor of performance. One objective of the support is to reduce dependency on the knowledge and on the experience of an individual financial advisor. Use of software is essential due to the complexity and the high data volume of this application domain. In PFP, the expected degree of individualization increases with client income and number of assets. Low income or few assets leads to standardization of the advisory service as the fixed costs have to be limited. With regard to the requirements of demanding customers and the technology-driven transformation process in the financial services sector, the acceptance of inflexible standard services decreases also in these customer segments.

4.1 Design and IT-Architecture

Life Charts is suited to develop a CBSS solution for decision making in personal financial planning combining the industrialization trends of the financial services sector and of software development. The component identification approach in [11] leads to a modular structured standardized advisory process for Life Charts according to the phases of PFP. The result of the approach includes the service elements as the identified components and the assembly schedule as the composition model. The bill of materials can be derived from the composition model, which is illustrated as an UML activity diagram based on black-box components in Fig. 3.

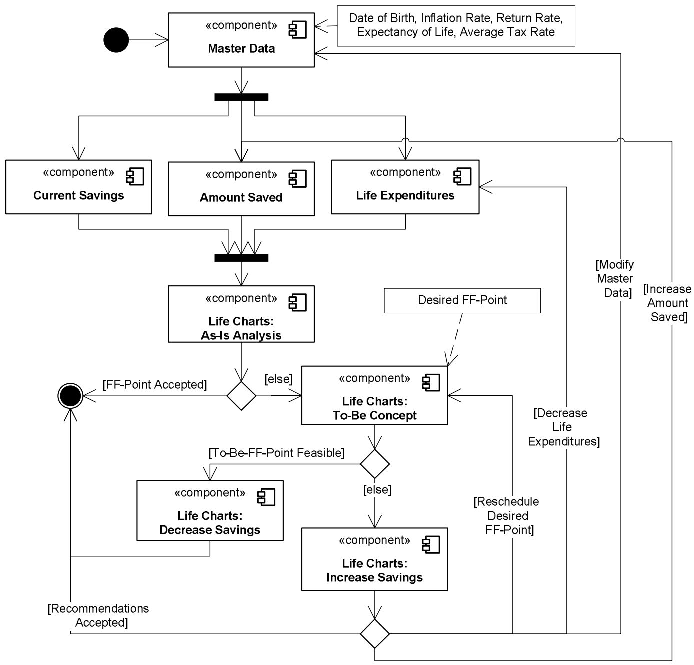


Fig. 3. Standardized Composition Model Life Charts

Following the design and implementation proposals of Section 3 the IT-architecture of the Life Charts CBSS solution (cf. Fig. 4) is based on SOA. Web service technologies and standards provide a basis for implementation. Components are mapped to web services in order to support the service elements. The process plans are transferred into a BPEL process composing the web services. The composition is not restricted to the given Life Charts services. External components can be integrated as well. An UDDI-module (Universal Description, Discovery and Integration) enables the retrieving of appropriate web services.

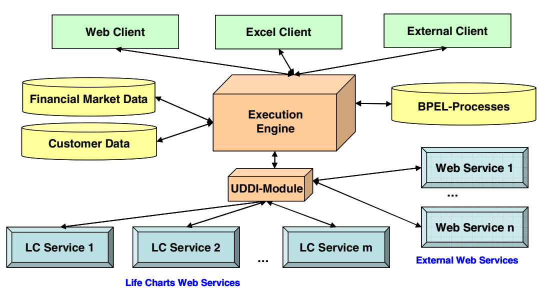


Fig. 4. IT-Architecture Life Charts

4.2 Prototype gsTools Life Charts

The standardized Life Charts process (cf. Fig. 3) answers the standard question (1) when an individual will reach the FF-point. Gathering *master data*, *current savings*, *amounts saved* and *life expenditures* provides the input for the equations (4.1) and (4.2). In the *as-is analysis*, the FF-point is calculated as the intersection of the functions of the existing savings $S(t)$ and the required savings $R(t)$. Variants of the standard question are answered too: If the calculated FF-point is not accepted by an individual, the *to-be concept* answers the question (1.1) how much required and existing savings are necessary at the desired FF-point. If the desired FF-point is feasible, the answer to the question (1.2) how much money can be spent more on life expenditures (*decrease savings*) is given. Otherwise proposals are made to the question (1.3) how *savings* can be *increased* to reach the FF-point earlier.

Fig. 5. BPEL Process Diagram Life Charts Action Planning

The visualization of the BPEL process implementing the action planning of Life Charts in terms of increasing and decreasing savings is shown in Fig. 5. Clients can execute the BPEL processes. Fig. 6 presents the screen of the Life Charts to-be concept of the web client prototype. The calculated FF-point of the as-is analysis (cf. Fig. 2) lies ahead of the desired FF-point. The trend line shows how the existing savings have to be increased in order to realize the desired FF-point. Thus proposals to increase savings are made next according to the BPEL process (branch on the left hand side of Fig. 5).

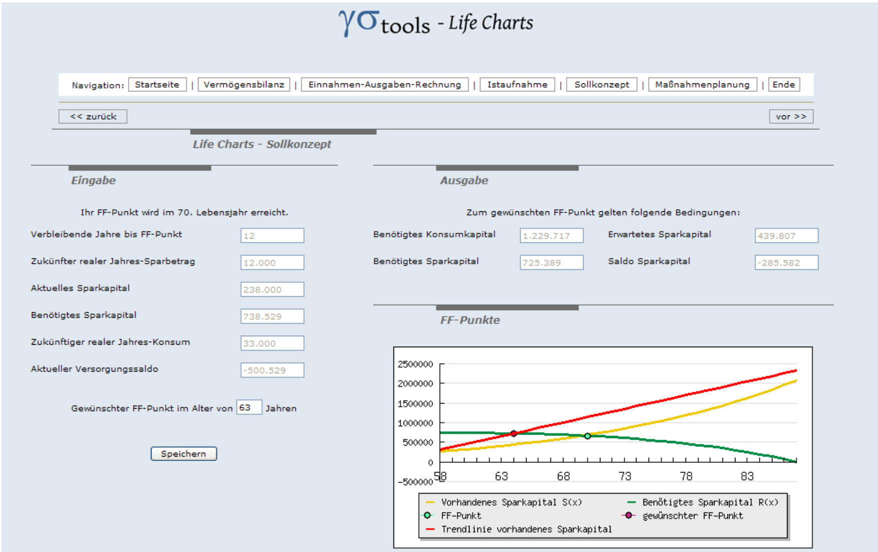


Fig. 6. Web Client Prototype Life Charts To-Be Concept

The standard advisory process just takes overall and constant assumptions about life expenditures, amounts saved and tax rates based on the current financial conditions. These assumptions may be adequate to average earners or entrants without predefined financial objectives. However, premium customers with detailed financial plans can neither specify irregular receipts and expenditures nor periods of life - specific income statements. For this purpose, a sophisticated *income statement* component is provided. This component enables to consider life events like purchasing a house and periods of life like sabbaticals, starting a family or retirement. These effects influence the FF-point calculation and increase the prediction performance of the answer to the standard question. In the same way, a *balance sheet* component states the current savings more precisely. The component *increase savings* can be replaced by a component providing concrete *recommendations for action planning* (cf. Fig. 7). Web services can be substituted in the implementation in an analogous manner.

Answers to further variants of the standard question can be found by adding new components to the composition model. The *recommendation* component can link to further components. Integrating a *portfolio optimization* component finds the answer to variant (1.4) of standard question (1) whether the desired FF-point is feasible by redeploying capital. The *risk management* component advises (1.5) how savings could be increased by old-age provisions in order to reach the desired FF-point. The *insurance comparison* provides the operational answer (1.6) which financial products are adequate to increase old-age provisions. Combining these components answers complex questions like (1.7) how much money have to be spent on old age provision to reach the desired FF-point when return on investment is already optimized by portfolio selection. The modified Life Charts composition model can be directly transformed into BPEL-process. Web services corresponding to the new components will get integrated.

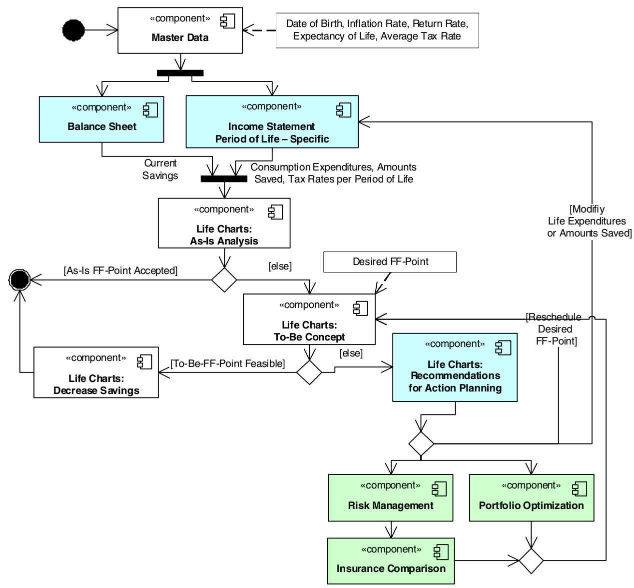


Fig. 7. Modified Composition Model Life Charts

4.3 Evaluation

The standard question of Life Charts is answered using the given components. Some variants of this standard question can be answered too. In order to answer further variants new components have to be integrated. In the Life Charts case study, each component not element in the given component set answers a question variant. Combinations of the integrated components answer further question variants.

We want to evaluate our approach also according to the criteria given in [12]. Capability of expressing complex compositions and quality related attributes like performance, development and maintenance effort are considered. Complex compositions are modelled in clear arranged variants of the process plans and are transferred to the BPEL processes. E.g. the answer to variant (1.7) of the standard question includes parallelism, decisions and loops. The suggested solution is compared to a reference implementation directly invoking PHP functions. The reaction time is at a slower rate. In the case of Life Charts action planning (cf. Fig. 5), the performance loss amounts to less than 0,1 seconds on average; using Tomcat 5.5, XAMPP 1.6.3a, AXIS and the ActiveBPEL engine 4.0 on a current desktop pc as server and a leased line. Encapsulating the functionality as web services means little development effort. PHP functions of the reference implementation are provided via a SOAP interface. The corresponding WSDL scripts of the web services have to be generated. Variants of the questions are answered by modifying the standard BPEL process which serves as a template. Existing web services and sub processes can be reused. Web services can be replaced by new ones matching the interfaces. No unexpected dependencies occur in the code and no changes in other web services are caused. Thus maintenance is simplified.

5 Conclusions and Future Work

The development of the prototype gsTools Life Charts applies industrialization to financial services and to information systems development. Service elements are mapped to components and are implemented as web services. Assembly schedules define the composition of the components. These assembly schedules are implemented as BPEL processes answering questions in the area of personal financial planning. This CBSS approach is evaluated with regard to the required number of components and the criterion of modeling complex variants. The number of components answering variants of the standard question of Life Charts is minimized. In Section 4.2, the standard question answers (1) when the point of financial freedom of an individual is reached. Within the given components the variants (1.1) - (1.3) of this standard question are answered. Each component added to the standardized composition model of Life Charts gives the answer to a further variant; cf. variants (1.4) - (1.6) in Section 4.2. Combinations of these components answer complex variants of the standard question; e.g. variant (1.7) in Section 4.2. The capability of substituting components enables MC, e.g. providing more precise answers for premium customers.

Quality related attributes of the proposed composition like performance, development, and maintenance costs are compared to a non-distributed approach of Life Charts. Performance is only marginally reduced. Additional effort related to providing functionalities as web services is limited and outweighed by the reduction of the remaining development effort: Development of product variants answering variants of the standard question is supported by simplified composition and reuse. Substitution and integration of components is facilitated by standardization and interoperability of the web service implementation. Maintenance effort is reduced by the assembly via defined interfaces and without unpredictable side effects.

The case study shows feasibility of our approach in the area of PFP. Additional applications and more quantitative evaluation are essential. Though composition of the web services in BPEL by hand is tool-supported, the support is limited; e.g. with regard to dynamic selection of web services. Syntax-oriented composition languages like BPEL are not sufficient to automate the composition (cf. [21]). Semantic annotations are required to enable reasoning. Therefore, we plan to compose semantic web services (cf. [22]) using OWL-S ([23]). The objective of the automation support is to reduce the manual composition effort and to facilitate the end users' involvement in the composition of business services processes.

References

1. Pine, B.J.: Mass Customization: The New Frontier in Business Competition. Harvard Business School Press, Boston, Massachusetts (1993)
2. Heiskala, M., Paloheimo, K.-S., Tiihonen, J.: Mass Customisation of Services: Benefits and Challenges of Configurable Services. In: Frontier of e-Business Research 2005 (FeBR 2005), Proceedings of the e-Business Research Forum conference (eBRF 2005), Tampere, Finland, 26.-28.9.2005, pp. 206–221 (2006)

3. Böhmann, T., Krcmar, H.: Modulare Servicearchitekturen. In: Bullinger, H.-J., Scheer, A.-W. (eds.) Service Engineering, Entwicklung und Gestaltung innovativer Dienstleistungen, Auflage, vol. 2, pp. 723–741. Springer, Berlin (2006)
4. Salvador, F., Forza, C.: Configuring products to address the customization-responsiveness squeeze: A survey of management issues and opportunities. International Journal of Production Economics 91(3), 273–291 (2004)
5. Sokolovsky, Z.: Industrialisierung der Banken. In: Sokolovsky, Z., Löschenkohl, S. (eds.) Handbuch Industrialisierung in der Finanzwirtschaft, pp. 33–58. Gabler, Wiesbaden (2005)
6. Moormann, J., Schmidt, G.: IT in der Finanzbranche. In: Management und Methoden, Springer, Berlin (2007)
7. Papazoglou, M.P., Traverso, P., Dustdar, S., Leymann, F., Krämer, B.J.: Service-Oriented Computing Research Roadmap. In: Dagstuhl Seminar Proceedings 05462, Service Oriented Computing (SOC 2006), pp. 1–29 (2006)
8. Szyperski, C., Gruntz, D., Murer, S.: Component Software – Beyond Object-Oriented Programming, 2nd edn. Addison-Wesley, Reading (2002)
9. Leymann, F., Roller, D.: Modeling business processes with BPEL4WS. In: Information Systems and e-Business Management, vol. 4(3), pp. 265–284. Springer, Berlin (2005)
10. Meta Group, White Paper in the context of OpenDoc (1994)
11. Olschok, B., Schmidt, G.: Identification of Configurable Components. In: Th. Blecker, K., Edwards, G., Friedrich, L., Hvam, F. (eds.) Innovative Processes and Products for Mass Customization, Series on Business Informatics and Application Systems, vol. 3, pp. 231–245. GITO, Berlin (2007)
12. Rabhi, F.A., Yu, H., Dabous, F.T., Wu, S.Y.: Information Systems and E-Business Management. A service-oriented architecture for financial business processes, Information Systems and E-Business Management 5(2), 185–200 (2007)
13. McIlroy, M.D.: Mass Produced Software Components. In: Naur, P., Randell, B. (eds.) Software Engineering: Report on a Conference by the NATO Science Committee, NATO Scientific Affairs Division, Brüssel, pp. 138–150 (1968)
14. Parnas, D.L.: On the Criteria To Be Used in Decomposing Systems into Modules. Communications of the ACM 15(12), 1053–1058 (1972)
15. Kilian-Kehr, R., Terzidis, O., Voelz, D.: Industrialisation of the Software Sector. Wirtschaftsinformatik 49 (2007) Sonderheft, S62-S71
16. Schmidt, G.: Prozeßmanagement, Modelle und Methoden, vol. 2. Springer, Berlin (2002)
17. Rosemann, M., van der Aalst, W.M.P.: A configurable reference modelling language. Information Systems 32, 1–23 (2007)
18. Papazoglou, M.P., van den Heuvel, W.-J.: Service Oriented Architectures: Approaches, Technologies and Research Issues. The VLDB Journal 16, 389–415 (Jul, 2007)
19. BEA Systems, IBM Corporation, Microsoft Corporation, SAP AG, Siebel Systems, Business Process Execution Language for Web Services visited: 23.07.2007 http://www.ibm.com/developerworks/webservices/library/ws-bpel
20. Schmidt, G.: Persönliche Finanzplanung, Modelle und Methoden des Financial Planning. Springer, Berlin (2006)
21. Mandell, D.J., McIlraith, S.A.: Adapting BPEL4WS for the Semantic Web: The Bottom-Up Approach to Web Service Interoperation. In: Fensel, D., Sycara, K.P., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870, pp. 227–241. Springer, Heidelberg (2003)
22. Bell, D., Bussler, C., Yang, J.: The Semantic Web and Web Services. Information Systems 31, 229–231 (2006)
23. W3C, OWL-S: Semantic Markup for Web Services, visited: 26.07.2007, http://www.w3.org/Submission/2004/SUBM-OWL-S-20041122/

Making the Agent Technology Ready for Web-Based Information Systems

Gheorghe Cosmin Silaghi

Babeş-Bolyai University,
Business Information Systems Department
Str. Theodor Mihali, No. 58-60, 400591, Cluj-Napoca, Romania
gsilaghi@econ.ubbcluj.ro

Abstract. The agent technology has become one of the most vibrant and fastest growing research area in Artificial Intelligence and Computer Science. But, there is still a big gap toward the wide scale adoption of the emerged agent technology by the software industry. Even if agent standards matured and stable agent platform implementations exist, the software industry still sees the adoption only after 2010 [1].

In this paper we propose a model for integration of agents in a web-based information system, from the standard information system engineering point of view. Rather than the agent-technology standards which sees the integration centered around the agent concepts, our approach is oriented around the standard object-oriented software process, which further allows the separation of roles in the software development team. Thus, average educated software engineers can approach agent solutions without the need of learning the agent technology, by only having the help of an agent expert. We exemplified our approach by presenting the design of a web-based application for information search, with the support of a multi-agent system.

Keywords: multi-agent systems, web-based information systems engineering, design patterns, software integration.

1 Introduction

Programming has progressed through machine code, assembly language, machine-independent programming languages, procedures and functions, abstract data types and objects to a new paradigm: agents. Intelligent Agents has become one of the most vibrant and fastest growing research area in both Artificial Intelligence and Computer Science [2]. From the software industry point of view, the question if the agent programming concept is valuable or not arises. Can we build large-scale agent systems for complex problems in large development teams? Or these models of increased intelligence are realizable only in small academic laboratories? How one can integrate the new agent technology with the existing software engineering tools? Are the agent technology tools mature enough to be reliable in

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 321–332, 2008.

real-live software systems? These are only few questions that one can ask regarding the emerged agent technology.

In the context of the actual transformations of the Web, making it intelligent with the support of the new agent technology supplied by the research community, becomes a need. The Web can bring the research efforts of creating intelligent software to the broad users and to the enterprises, if a proper integration is possible between the "intelligent software" and the web applications.

The objective of this paper is to investigate if agent technology is mature enough to be adopted by the software industry, in the sense that enough software practices exist for a rapid deployment of a multi-agent system as part of a common (web-based) business information system. Agent research and their FIPA standards [3] envisage the software development centered around the agent technology. Therefore, when designing an agent system for a business problem the software engineer needs to think, work and reason with agent concepts which, in most of the cases, they are not familiar with. More, FIPA standards see the interface between the multi-agent systems and other existing information systems by embedding the latter ones as input end-points in the multi-agent systems. This approach is very complicated for a average-educated system engineer who, in the majority of cases, for a particular problem, will rather drop the agent technology-based solution and consider instead a standard software engineering solution. Although there were big expectations and promises, the agent technology did not succeed to penetrate the industry as predicted [1] and a reason for this might be the after-mentioned issues.

In this paper we propose a different approach for the engineering of agent systems inside standard web-based information systems. Instead of requesting software engineers to learn agent-oriented concepts and software process, we will let them to continue to work and reason with their basic (mostly) object-oriented tools and plug-in the multi-agent module as a new service inside the engineered information system. An agent expert might become responsible for designing and developing the multi-agent system. Using the guiding example we describe in this paper, the software engineer would be able to integrate the agent system inside the new information system under development.

To exemplify our approach, this paper will describe the deployment of a multi-agent system for web search inside a web-based information system, engineered with the aid of the common-employed Master View Controller design pattern [4] for J2EE technologies.

The paper is organized as follows. In section 2, we describe the actual context, presenting the background on which we build our contribution. Section 3 presents the business problem definition to be solved with agents and the initial design of the agent solution (AgentSearch). Section 4, present our technical solution regarding the topic integration of the agent platform into the website, while section 5 exemplifies this solution on AgentSearch. Finally, section 6 concludes the paper.

2 Context

2.1 The Classical MVC Design for a Web-Based Information System

Software development is an activity that requires a high level collaboration inside the team that approach the development phases: the business analysts, the software developers, database and conceptual designers and testers. To bring productivity inside the development team, modularization of the software process and the software engineering phases is a must. With this respect, design patterns [5] came in front, bringing the most wanted development productivity, by the re-usage in large scale of existing solutions and templates that proved to work well on small-sized specific problems.

As defined by Shalloway & Trott [5], the Model-View-Controller (MVC) pattern is primarily used when building graphical interfaces and can be used anytime in an interactive system. It is used to decouple the data, the presentation of the data and the logic for handling the events from each to other.

Most web-tier application frameworks use some variation of the MVC design pattern [4]. Fig. 1 describes the MVC architecture for a Java web application. This architecture leads for a horizontal decomposition of the web information system in layers, each layer having a clear role in the web application and being composed by independent smaller details sub-components. Software design and programming at each layer level is realized with extensive usage of basic design patterns [5], like singleton, abstract factory, iterators, proxies etc. Fig. 2 presents the standard horizontal layered architecture for a web information system.

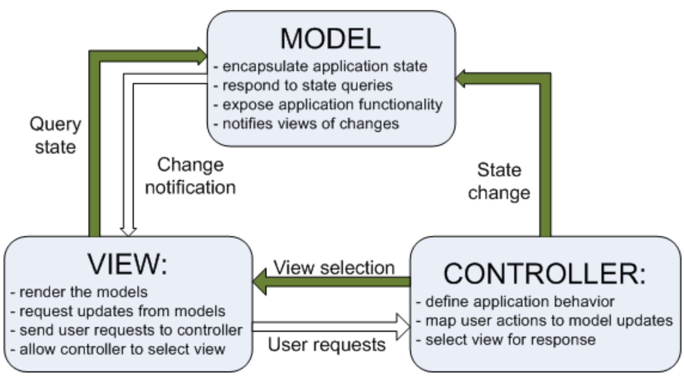


Fig. 1. The MVC architecture for a Java web application

2.2 Agent Systems and Integration with Other Software Services

According to Wooldridge [6], an agent is a computer system that is capable of independent (autonomous) action on behalf of its user or owner, figuring out what needs to be done to satisfy the design objectives, rather than constantly

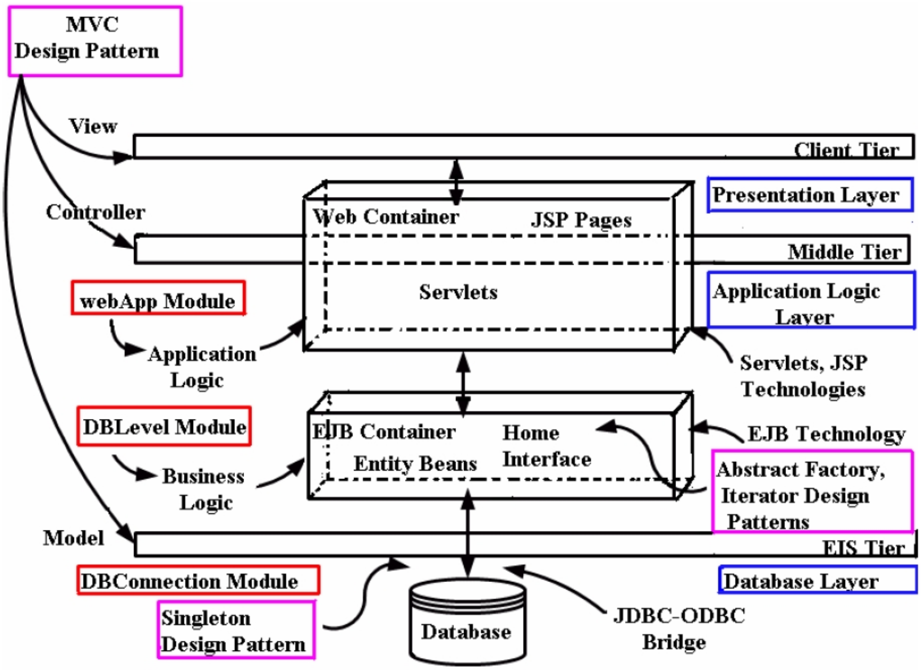


Fig. 2. The horizontal layered decomposition of the standard MVC architecture for a Java web application

being told. Up-to-date, a lot of research was performed, leading to the maturation of the agent technologies and even to standardization.

Agent standardization activities started from 1998, when the first usable FIPA standards [3] were developed, These standards became mature only on 2000 and AgentLink [7] recommended their usage for the large scale development of multi-agent systems. The recognition of the agent field and the importance of FIPA standards for practical agent development came in the June 2005, as FIPA, the standards organization for agents and multi-agent systems was officially accepted by the IEEE as its eleventh standards committee.

Regarding the software design methodology, FIPA describes its artifacts using UML and object-oriented tools, but does not state anything about some recommended software engineering methodology. Up-to-now, several FIPA-compliant platforms are available (like ZEUS [8], JADE [9] , FIPA-OS [10]), JADE being among the most matured and preferred ones by the research community and the industrial partners. We should note that Java-based implementations for multi-agent platforms are predominant in the agent community.

Regarding the interoperability issues of agents with external software systems - here of interest, FIPA Agent Software Integration specification [11] describes how agents have to broker and negotiate over software systems and how new software services are to be dynamically introduced into an agent community. This

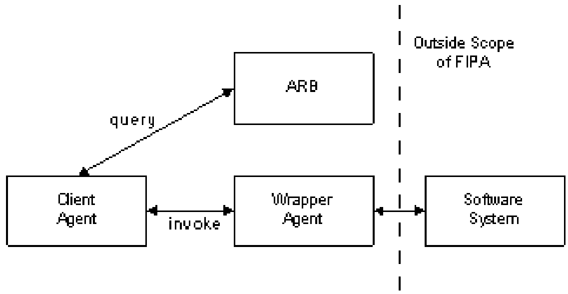


Fig. 3. FIPA General agent software integration scenario

specification enables developers to build *wrappers* for software services which are to be utilized and/or controlled by a community of agents, thus making agents ready to access the "public" services offered by external information sources. Fig. 3 represents the general agent software integration scenario. Therefore, a FIPA client agent may query an Agent Resource Broker (ARB) about available software services on the platform. Software services are wrapped inside a wrapper agent for FIPA compliance. In order to use such a service, the client agent will deal in a FIPA-compliant communication act with the wrapper agent.

This agent-centered approach was pursued in the implementation of the agent-based travel support system of Gawinecki *et al.* [12,13]. They broke the system into functionalities accomplished by different agents, the web-interface logic being the responsibility of some proxy agents. Agents encapsulate everything, including the front-end logic and session handling, they acting as servers for the web-based information system. We differentiate from this approach by letting the standard functionality of the front end to be designed in a classical way, and putting the agents to work only on specific issues.

As the AgentLink roadmap reveals [1], although the agent technology matured and get accepted by the technical community through standardization, there is still a big gap till the the wide adoption of agents by the industry. On a deliberate Delphi survey including 18 experts from the industry, fields like manufacturing, transport, telecommunications, networks and healthcare see the adoption of agent technology in the next 5 years (starting from 2005), while computer software people are more reluctant, seeing the adoption in a frame of 10 years. Mainstream deployment of agent technologies will not appear until 2010. Probably one reason for this conclusion could be the fact that software industry people are not prepared enough for the new technology and there is always a reluctance from the enterprises to switch the existing patterns of development for novel and less-experienced solutions.

3 AgentSearch - Agent-Based Web Information Retrieval

In this section we present the definition of a web-search problem to be solved with the help of agents. AgentSearch - the system we deployed to solve the search problem, represents a validation scenario for the technique we propose regarding the integration of agents into web applications. Therefore, we will not insist on internal design issues of AgentSearch and of the incorporated technologies.

3.1 Problem Definition and Requirements for the Agents

Within the context of the growth and the wide usage of Internet in all day life, the question about how to find the relevant information on web arises. As we are focusing on agent technology, our target is to build intelligent information agents which are able (1) to access multiple, distributed and heterogeneous data, information and knowledge sources, (2) to acquire, store, mediate and maintain the relevant information with some respect, (3) to behave at the request of some users or a particular user which may be human or another agent. Technically, our agents will be able to acquire the information by crawling the web, find the relevant information for a query on the database of documents and cooperate with other agents in order to exchange knowledge based on the other's expertise.

To perform these tasks the agents need to use techniques from (1) information retrieval and link-directed web search for acquiring are retrieving relevant information; (2) social filtering and content-based recommendation for personalizing the retrieved information according with their user or usage; (3) data mining, information extraction and summarization techniques for tuning the information or getting into their insight. We will not enter the details of these techniques, as they are beyond the scope of this paper. We only want to emphasize that they are complex enough to justify the usage of a multi-agent system, instead of a standard (non-intelligent) information system. In order to put the agents to work on behalf of the users, we envisage the use of agents through a web interface, thus we end up building a web-based information system.

3.2 Multi-agent System Design Considerations

In this subsection we will shortly describe the design of our multi-agent system, named AgentSearch. In this system we designed two sorts of agents:

- *domain agents*: responsible for carrying out indexing tasks. Indexing tasks may occur in the following situations: (1) at the very beginning of the execution of the system, when no documents are indexed; (2) at some request from a user agent. When a user agent feels that its user is not happy with the recommendations she receives, it might ask the a domain agent to start expanding its current neighborhood graph; (3) at some external request, when all domain agents should proceed with re-doing the indexing, in order to keep the document database in a consistent form
- *user agents*: who search information for the behalf of their users. User agents retrieve pages for the query, compute the relevance measures of those pages and collaborate with other user agents in order to improve the result set by collaborative filtering. User agents observes how their users use the retrieved information and maintain a user profile. User agents might interact with domain agents in the situation when retrieved information for a query scored a low relevance.

The main use case of AgentSearch has the following steps:

- the user comes to the system and specify some domains of interest. Each domain is in fact an URL representative for that domain
- the user agent responsible for that user invoke domain agents to crawl the web starting from the initial domain URLs and build a representative documents database for the user
- the user enters queries into the system
- for each query, the responsible user agent searches first the representative documents database for the user. Next, the user agent asks other user agents to retrieve documents from their (partial) document database. In the end, the responsible user agent merges the recommendations received from various other user agents, ranks them according also with collaborative filtering principles (i.e. similitude between users, based on the user profiles) and produces the recommendations to the user
- the user submits a feedback to her user agent, regarding how satisfied she was from consulting the recommendations
- the responsible user agent maintains a user profile, incorporating all the received feedback

Fig. 4 shows all the subcomponents of an agent and the logical interconnection of those subcomponents. We should note that while the "Web indexer and document processor" and the "domain manager" components are specific for the domain agents, the rest of components are specific for the user agents. This figure also shows the external interaction an agent might have, with the graphical

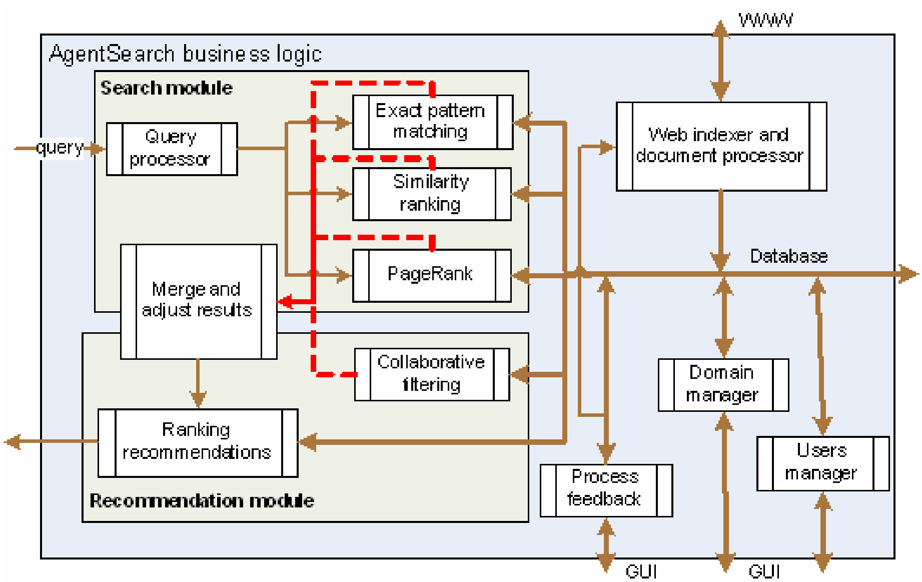


Fig. 4. AgentSearch: decomposition of the business logic at the agent level

interface of the information system, with the underlying database and with the web.

4 Integrating the Agent System in the Web Application

In this section we present our approach for integrating a multi-agent system in a classical web application architecture.

The model presented in fig. 3 is the one we intend to improve as part of this paper. If agent software development is to be done according with this model, it means that the agent model will be the one adopted for the overall solution of an information system. That means the software engineer and the development team behind him should be aware of this model and should know the details of agent technology in general, of FIPA specifications and its implementation in particular. These assumptions usually do not happen and, as a consequence, it is likely that a software engineer will not adopt a multi-agent solution, even if such a solution might be the best one for some particular problem.

Instead, starting from the wide-accepted and used MVC layered architecture presented in fig. 2, we propose to integrate the multi-agent platform as a distinct component(the agent service layer), part of the application logic layer. All the multi-agent system needs to expose is its service layer that will collect all the interaction of the multi-agent system with the controller of the web application. Thus, the web application will continue to function as it was previously designed. The fact that the application logic encapsulates a complicated and intelligent structure like a multi-agent system will be transparent for the web application. The concept of the agent service layer is very similar with the well-known web services, that hide classical applications or information systems behind some web-identifiable end-points.

We should note that the agent service layer actually exists in the FIPA standards, the *service* concept being a mandatory and central element in the FIPA abstract agent system architecture [14], with the following short description: "a service provided for agents and other services". Thus, FIPA envisages that agents will supply services between them, these services will be registered in the service directory and agents will use ontological support in order to learn to communicate with other agents services.

We should emphasize that the agent service layer needs to keep inside only the intelligence of the system and should not encapsulate standard functionality like sessions or navigation control of the website, which are very well realized by standard tools (like Apache Struts [15]) implementing the MVC model.

In particular, JADE, being a FIPA compliant platform implements the service layer. Each JADE agent is an instance of the class `jade.core.Agent`. The programmer can control the agent through a handler which is of the predefined class `jade.wrapper.AgentController`. Therefore, what the programmer needs to do, is to implement a new class (we named it `AgentService`) representing the service layer that composes objects of type `AgentController` used to send specific messages (and invoke agent services) to the agents. Specific functionalities that the agent system exposes to the web application are represented as end-point functions in the `AgentService` class. Such a function in fact transfers the information flow from the web application toward the agent platform, including the conversion of the messages in the languages understood by the agents. The `AgentService` class should also contain a reference for the agent platform. If the agent platform is customary for the our web application, then, when instantiating the `AgentService` object we can create the agent platform. If the agent platform exists outside the scope of our information system, then the reference for the agent platform can be supplied as an input for the creation of the agent service layer. The reference to the agent platform is represented as the composition with the `jade.wrapper.ContainerController` class. Like other components in the web application, is it recommended that the `AgentService` class to be implemented as a singleton, therefore, assuring a sole entry point for the multi-agent system. Fig. 5 represents the implementation of the agent service class for the sake of our example. Algorithm 1 represents the sketch of the source code.

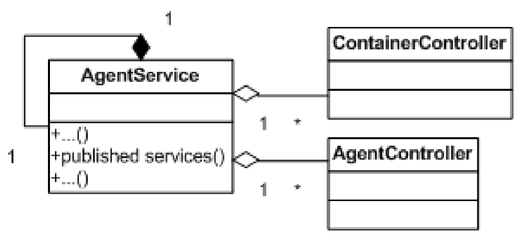


Fig. 5. The MVC architecture for a Java web application

The technical solution we presented in the paragraph above is a very simple one, from the programmer's point of view. The programmer creates similar classes, e.g. when interacting with database layer. Therefore, in this way we can decouple the job of the software engineer and of the programmer from the much more specific job of the agent engineer. The development team can now be composed of members with different roles (like the agent engineer), without requiring too much overlapping of their responsibilities.

5 Practical Integration of Agents in AgentSearch

In this section we present the practical design of AgentSearch, showing how the multi-agent system fitted within the MVC architecture of the website. As part of the MVC design, we used the following productivity tools: iBATIS [16] for separating the SQL code in the database layer from the JAVA programming language used in the software project and Apache Struts [15] for managing the application flow at the application logic layer, inside the web container. Apache Struts is built specifically for the MVC. Jade was used as agent platform.

Algorithm 1. Source code implementing the AgentService class

```
class AgentService {
  static private AgentService mAgentService = null;
  private ContainerController mAgentMainContainer;
  private AgentController rma = null; // remote management admin
  // ... other class members
static public AgentService getInstance() { // singleton
  if (mAgentService==null) {
   mAgentService = new AgentService(); }
  return mAgentService; }
private void InitializePlatform() {
  ...
  rma = mAgentMainContainer.createNewAgent("rma","jade.tools.rma.rma",null);
  rma.start(); }
private AgentService() { // create the main container and initialize the platform
  Runtime rt = Runtime.instance();
  Profile p = new ProfileImpl();
  mAgentMainContainer = rt.createMainContainer(p);
  InitializePlatform(); }
} // end the class definition
```

Fig. 6 shows the layered design of the web-based information system that accommodates AgentSearch. This picture instantiate all theoretical layers described in fig. 2 and includes the multi-agent system as a distinct layer. We should note that the interconnection between the controller and the agent platform is realized through a specific agent service layer, composed by the `AgentService` class described in fig. 5. The web layer depicted of the left side of the picture at the same level with the JADE Agent Layer is in fact connected and coordinated by the domain agents.

As fig. 6 shows, the leveled integration of the multi-agent system keeps the old design principles fulfilled. More, as in this system the agent platform represents the core application logic, the agent platform simply one-to-one substitutes the standard application logic depicted in fig. 2. The decomposition of the web information system is natural, thus, not requiring further innovations or the design of unusual programmable fabrics. More, there is a natural delimitation between the multi-agent platform and the rest of the system, the interaction and control being filtered only through the agent service layer. Therefore, we can employ an agent specialist in the development team and assign him a crystal-clear development role, keeping all the rest management of the software process and development team unchanged.

We consider that this pattern of development can be followed every time when intelligence is requested in a information system. As majority of information systems built today are organized around object-oriented software engineering principles, our approach and development experience shows a smooth integration of intelligence techniques. We recommend the re-production of these development

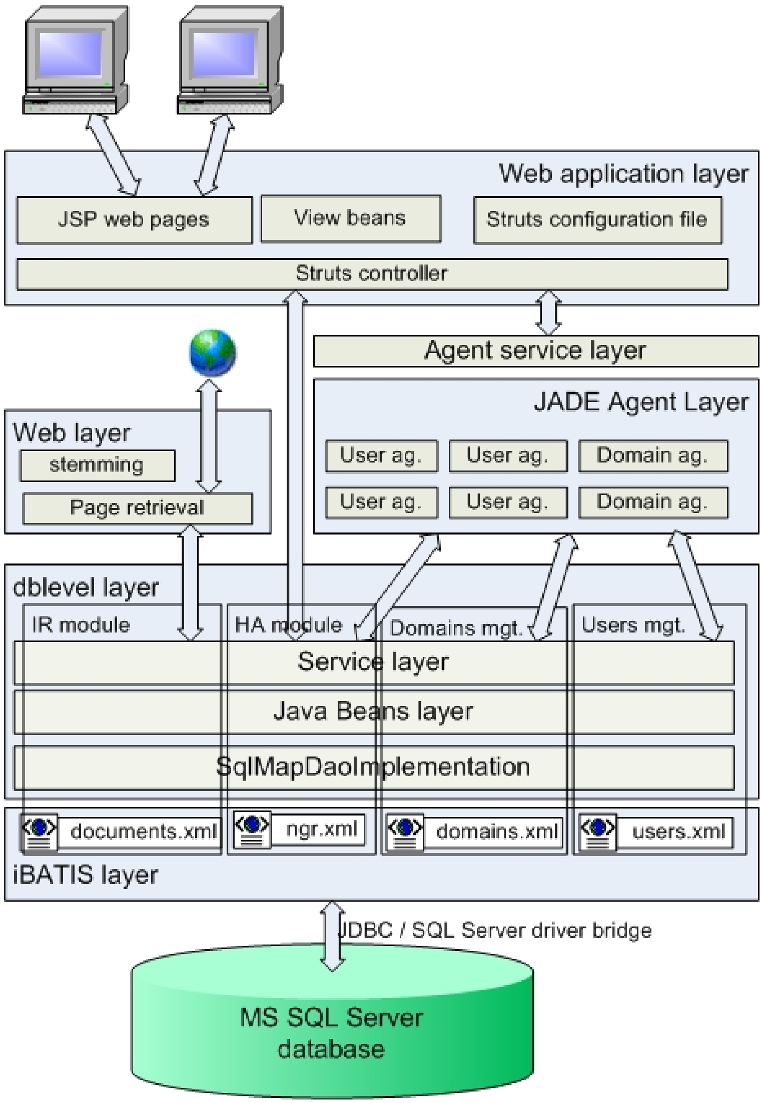


Fig. 6. AgentSearch: decomposition of the business logic at the agent level

experience every time when one team faces the contact or needs the input of the multi-agent technology.

6 Conclusion

In this paper we investigated the easiness of integrating agent systems in a classical web-based information system. While the agent research community sees this integration centered around the multi-agent system, this might represent a difficulty for an average-educated software engineer to adopt the agent-based solution for a particular problem. Thus, our contribution is to suggest an engineering pattern to be followed in order to smoothly integrate agents in common (web-based) information systems. Our integration scenario has the advantage that is centered around the information system to be built and decouples the agent platform from the rest of the system. This allows the specialization of the system development team members, thus, integrating and agent engineer inside team and leveraging the rest of the team members from knowing or learning the details of the new agent technology.

We exemplified our integration solution by showing the design of a J2EE MVC-centered web-based application for information search, with the help of FIPA-compliant agents for improved efficiency.

Acknowledgments. This work is funded by the Romanian Authority for Scientific Research under the project IDEI_573 and by Babeș-Bolyai University Cluj-Napoca, Romania, under the TP_T project no. 30948/12.07.2007.

References

1. Luck, M., McBurney, P., Shehory, O., Willmott, S.: Agent Technology: Computing as Interaction. In: A Roadmap for Agent Based Computing, Agentlink (2005)
2. Mangina, B.: Review of software products for multi-agent systems. Agentlink (2002)
3. The foundation for intelligent physical agents, http://www.fipa.org
4. Singh, I., Stearns, B., Johnson, M.: Designing Enterprise Applications with the J2EE Platform, 2nd edn. Prentice-Hall, Englewood Cliffs (2002)
5. Shalloway, A., Trott, J.: Design Patterns Explained: A New Perspective on Object-Oriented Design. Addison-Wesley, Reading (2001)
6. Wooldridge, M.: Introduction to MultiAgent Systems. Wiley, Chichester (2002)
7. Agentlink: European co-ordination action for agent-based computing, http://www.agentlink.org
8. Zeus agent toolkit, http://sourceforge.net/projects/zeusagent
9. Java agent development framework, http://jade.tilab.com
10. Fipa-os agent toolkit, http://sourceforge.net/projects/fipa-os
11. FIPA: Fipa agent software integration specification. standard no. XC00079B (2001), http://www.fipa.org/specs/fipa00079/.
12. Gawinecki, M., Gordon, M., Kaczmarek, P., Paprzycki, M.: The problem of agent-client communication on the internet. Scalable Computing: Practice and Experience 6(1), 111–123 (2005)
13. Gawinecki, M., Kruszyk, M., Paprzycki, M., Ganzha, M.: Pitfalls of Agent System Development on the Basis of a Travel Support System. In: Abramowicz, W. (ed.) BIS 2007. LNCS, vol. 4439, pp. 488–499. Springer, Heidelberg (2007)
14. FIPA: Fipa abstract architecture specification. In: standard no. SC00001L (2002) http://www.fipa.org/specs/fipa00001/b
15. Apache struts framework, http://struts.apache.org
16. ibatis data mapper framework, http://ibatis.apache.org

An Architecture for a QoS-Aware Application Integration Middleware

Daniel Martin[1], David de Francisco[2], Reto Krummenacher[3], Hans Moritsch[4], and Daniel Wutke[1]

[1] University of Stuttgart
{daniel.martin,daniel.wutke}@iaas.uni-stuttgart.de
[2] Telefónica Research and Development
davidfr@tid.es
[3] STI, University of Innsbruck
reto.krummenacher@sti-innsbruck.at
[4] Vienna University of Technology
hans@complang.tuwien.ac.at

Abstract. Complex application integration scenarios often demand for different combinations of qualities of services (QoS) at middleware level. The scenario presented in the paper, for instance, requires transactional middleware behaviour for business negotiations between a relatively small number of participants on the one hand, while on the other hand it requires high scalability for distributing data to a large number of clients. The concept of *Triple Space*, a semantically enhanced, distributed tuplespace middleware based on an extended Linda model, has been developed to provide such an infrastructure. In contrast to existing middleware infrastructures, Triple Space supports a set of configurations which define the system's QoS. In this paper, we present a motivating use case scenario, deduce requirements for the architecture of Triple Space, define its architecture and three QoS configurations, and outline our approach towards implementing a highly scalable distributed communication infrastructure.[1]

Keywords: Scalable Tuplespaces, Triple Space, Semantic Web, QoS.

1 Introduction

The World Wide Web as it exists today provides a platform for human interaction through sharing of human-interpretable content using the paradigm of *persistently publish and read* [4]. In contrast, Web Services enable machine-based interaction and follow a messaging oriented approach comparable to e-mail. The TripCom[2] project aims to introduce Web services based on the principle of *persistent publication*, enabling Web-like communication for Web services. This is

[1] This work is funded by the European Commission under the project TripCom (IST-4-027324-STP).

[2] http://www.tripcom.org

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 333–344, 2008.

accomplished using tuplespaces [8] as the underlying technology to support persistent communication and strong decoupling in the dimensions of reference, time and space [4]. Furthermore, TripCom leverages Semantic Web technologies to enable machine-processable data in the form of RDF triples to be communicated over a semantic-aware tuplespace - the *Triple Space*. Triple Space is a global communication and coordination infrastructure based on writing, observing, and retrieving distributed semantic data. TripCom based services are therefore also decoupled in the dimension of data structure, allowing the resolution of schema heterogeneities through mediation. We implement Triple Space as a distributed system without a central, single point of control. While it conceptually represents one global space, it is physically distributed over a possibly large set of interconnected kernels.

Being able to scale-out is a big challenge for today's IT systems, especially when parts of the business are directly connected to end-users over the Internet e.g. through web portals. In these scenarios it is for instance a typical requirement that the infrastructure must be able to cope with extreme peaks of workloads for short periods of time. Web middleware typically can deal with these kinds of situations, it does not however provide the quality of services (QoS) known from enterprise middleware systems, especially from the area of enterprise application integration (EAI). These kinds of middleware systems are build to provide quality of services (QoS) required by business transactions, e.g. it must be guaranteed that data is always in a consistent state or that messages must not be lost under any circumstances. Systems must be able to recover gracefully after failure, even if only parts of the underlying infrastructure are available. Requirements for Internet-enabled systems and for intra-enterprise systems therefore can be disparate and sometimes even mutually exclusive: a highly scalable system buys its scalability at the expense of not being able to provide any guarantees on data consistency or completeness of search results. Whereas a system that provides ACID transactions does not scale due to the huge amount of communication necessary to provide data coherence and consistency.

The dilemma however is – with today's ubiquity of the web – an enterprise system sometimes needs to provide both ends of this spectrum at the same time. We approach this problem by introducing so called *sub-spaces* that are able to provide QoS known from traditional EAI middleware. This is done by assigning so called *configurations* – precise descriptions of necessary QoS – to sub-spaces. It is then possible to integrate data from these individual *islands of strong QoS* in a large scale infrastructure (the *space*), that does not have strong QoS, but is able to scale-out and provide simple to use access methods to reach the data residing in sub-spaces.

This paper is organized as follows: in Section 2 we present two real-world use cases as an example of the aforementioned scalability vs. strong QoS dilemma. Subsequently, we describe configurations we identified, followed by an architecture of a Triple Space system that implements them. The logical architecture of one Triple Space kernel and a description of its components is presented in Section 3.1. Based on this work, Section 3.4 describes our approach to achieve high scalability through distribution by focusing on the architecture of inter-kernel communication. Section 5 concludes our work and gives an outlook on future directions.

2 Use Case

Current business trends orient service offerings to collaborative scenarios, with a fine-grained value chain and specialization of partners to provide added-value services. The collaborative nature of these scenarios makes integration a critical requirement, derived from the heterogeneity of the data sources which needs to be tackled. Many collaborative scenarios follow a marketplace like communication [2]. We define a marketplace as a generic business model, where multiple sellers and buyers compete (many to many interaction) in an open and ad-hoc manner to buy and sell goods. Buyers and sellers meet in a common negotiation space to perform business transactions in an open and dynamic environment. Orders can come from both sellers (e.g. offer goods) and buyers (e.g. request goods), and are subject to speculative strategies influenced by market conditions and individual strategies of the interacting parties. A typical example of a marketplace like communication between actors is the stock exchange market, where actors exchange stocks through business transactions.

2.1 Digital Content Negotiation

In previous work [7], business relevance of Digital Asset Management for telecommunication industry was argued. Telefónica, as a telecommunication operator and owner of an extensive communication network infrastructure, offers a television over IP service called *Imagenio*[3]. As service provider in this scenario, Telefónica does not generate its own content but rather receives it from content providers, which generate content (i.e: film producers, TV companies). To accelerate the process of offering content to end customers, Telefónica decides to interact with its suppliers in an open and ad-hoc environment: an auction model (a specialization of the marketplace model [2]) is implemented by Triple Space (see (1) in Figure 1). To accomplish the business transactions, the following functional requirements, among others, are needed: (i) Telefónica, as the auction controller, needs to be able to dynamically organize and control the auctions. It is responsible for creating sub-spaces in the company-wide triplespace and assigning required QoS by means of "configurations" (see Section 3.1). (ii) The auction system's objective is to perform business transactions. The information exchanged during business transactions must be obviously consistent. With this aim, the usage of transactions to perform these kind of negotiations

[3] The service is real and being offered at this moment (see `http://www.imagenio.com/`). However, negotiations are not carried out in the way we describe in this section. This example however is a realistic business model feasible with today's infrastructure and might be considered by Telefónica as a future implementation.

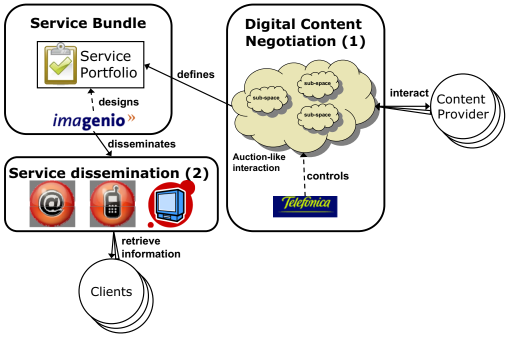


Fig. 1. Marketplace Interaction Example: Digital Content Services

in an atomic and secure way is critical to the system. In other words, business transactions must be either finished successfully or rolled back in case an error occurred and the transaction failed. The number of partners within the negotiation scenario are somehow limited, since only companies trusted by Telefónica would be invited to participate, thus the number of interacting partners can be predicted – or at least an upper bound can be given.

2.2 Large Scale Information Distribution

Continuing with previous example, data from the negotiation leads to the composition of digital content services offered by Imagenio, which will be offered to its customers via Web, mobile and TV interfaces (see (2) in Figure 1). The objective in this use-case is the dissemination of available contents of Imagenio service to all potential customers, using as many communication channels as possible. In contrast to the use-case described above, this use-case does not demand transactional behavior, but instead has strong requirements on systems scalability, since the number of potential customers that might be interested in content Imagenio is offering at a given point in time can not be predicted. In addition, the following functional requirements have to be fulfilled: (i) the publication of the information in a neutral format is needed – exactly what semantics can provide by the definition of ontologies as common vocabulary for all the partners involved in the scenario and the transmission of semantic data. (ii) A mechanism to retrieve this information in a simple manner is needed. Space based computing, by using template matching mechanisms allows users to retrieve required information and to coordinate individual pieces of application logic in an ad hoc and loosely coupled manner.

The scenarios described above have completely different requirements on the communication middleware, but share common data. A traditional approach would be to extract data from the system that implements the first scenario, then load it into the system that implements the second scenario. A major advantage of TripCom in this example is that it supports these two distinct scenarios despite their disparate requirements. The additional step of extracting data from one system and loading it into another is not necessary any more, allowing to provide the final service to customers and thus to run the whole process in a much faster and convenient way.

2.3 Benefits of Triple Space

Business processes as outlined in previous sections typically comprise several different applications hosted by multiple partners that were deployed individually and are mostly designed without integration in mind. Such applications are naturally heterogeneous, autonomous, distributed, and immutable; they have their own data and process models, are designed to run independently, operate on local data stores and have limited adaptability to the overall IT infrastructure. This problem is widely known and the very reason for the existence of EAI middleware, which is typically based around message-oriented middleware (MOM) plus application adapters and routers [9] to implement the required flow of information and data. Today's EAI middleware consists of a set of products designed for specific aspects of the whole problem, namely: *suitable communication infrastructure*, *common message format and protocols*, and *agreed-upon data semantics*. Provided QoS are typically related to the delivery of messages; systems for instance ensure guaranteed delivery by employing store-and-forward routing between individual nodes which are connected in a broker or bus-style manner. The key benefit of this style is the introduction of a target application-independent, neutral message format (e.g. EDIFACT) to reduce the number of required message transformations from $O(n^2)$ as it is with point-to-point integration of each participating application to $O(n)$. TripCom's interaction model provides enterprise integration with both, the ability to coordinate multiple applications and reliable communication at the same time. This is realised by writing data and information to and retrieving it from a shared information space (cf. persistent publication and read [4]). Applications observe shared data through blocking operations or asynchronously through notifications and can react to state changes according to the business process they are part of.

Ontologies, controlled vocabularies that formally describe a business domain in terms of domain-specific concepts and the relations between them, provide a widely-accepted instrument to cope with the problem of lack of semantics in business messages [3]. References to commonly agreed ontological concepts can be transmitted alongside the message payload, enriching the semantics of the original textual descriptions. If multiple ontologies for the same domain are available, their joint usage can be achieved via user-defined mediators, in form of formal ontology mapping languages or custom code [11]. With the help of ontologies and ontology mapping methods, an EAI system is capable of automatically transforming between different representations of data with the same meaning. TripCom was designed from the ground up to support ontologies and mediation. Its basic data format is RDF and it supports mediation in the form of external services or predefined transformation rules (see Section 3.1). It uses triple stores that also provide reasoning capabilities for the implementation of its persistent storage layer. Not only was TripCom designed for typical EAI scenarios as outlined before, it also provides configurable qualities of services (QoS) that allows the infrastructure to be used in disparate scenarios that usually require totally different types of middlewares due to the mutual influence of scalability and consistency. To solve this problem, the concept of *configuration* is introduced in the next section together with an description of a supporting architecture.

3 Triple Space Architecture

To enable users to configure their communication infrastructure according to use case scenario requirements, the notion of *configurations* has been introduced in Triple Space. Configurations combine (i) functional characteristics of Triple Space (i.e. supported operations) and (ii) non-functional properties (e.g. transactions, search completeness). So far, three configurations have been defined, covering the full spectrum between Web scale and database-like QoS.

The scenario for a configuration 1 Triple Space is a globally distributed network of a large number of Triple Space *kernels* (the architecture of a Triple Space kernel is presented in Section 3.1) that are inter-connected via a network and independently managed by separate authorities. A configuration 1 Triple Space aims for scalability similar to the Web by restricting functionality to a minimum and allowing for the most liberal QoS. Data access is restricted to individual RDF triples rather than graphs, the query language is restricted to simple template matching and operations for data removal are omitted. With regard to non-functional characteristics, a configuration 1 Triple Space does e.g. not guarantee data consistency or search completeness. However, even with the aforementioned restrictions, Triple Space goes beyond the functionality provided by the Web since it enables content-based data access – data can be retrieved by a template rather than by reference in form of a URL. A client gains access to the global triplespace by connecting to an arbitrary kernel. The kernel exposes the Triple Space API and takes over routing of requests to other kernels. Thus, a client does not have to know the particular kernels that provide the requested services directly. The global triplespace consists of a potentially large number of triplespace kernels. Each kernel is responsible for (i) providing access to the triplespace, (ii) coordinating connected clients [8], (iii) retrieving required data from other kernels, (iv) mediating between different data formats, and (v) guaranteeing secure data exchange.

Triple Space configuration 2 extends configuration 1 on a functional level by adding atomic bulk operations for triple read and write and providing an RDF query language similar to SPARQL [16]. Therefore, operations are not only possible on single triple level but also on the level of RDF graphs consisting of multiple interrelated triples.

Triple Space configuration 3 extends configuration 2 with functionality for removal of triples and graphs. Furthermore, this configuration provides full support for ACID transactions. This configuration also allows clients to register

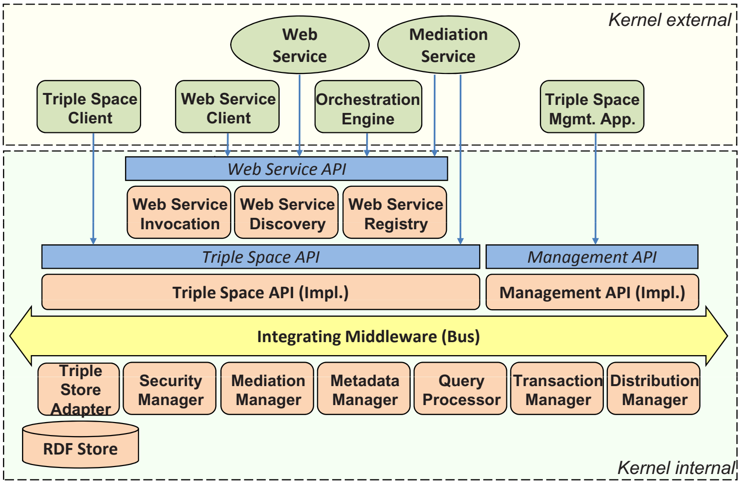


Fig. 2. Triple Space Kernel Architecture

(SPARQL-) templates together will a callback in order to be notified if matching triples/graphs are inserted into the space.

3.1 Kernel Architecture

The functional building blocks of a triplespace kernel are reflected in a number of kernel components which are implemented as pluggable, autonomous modules communicating with each other over a kernel-internal bus system. This approach to component integration provides for a scalable kernel implementation as it enables (i) connecting multiple instances of a particular component to the middleware (and therefore the whole triplespace kernel) to prevent a component from becoming the system's bottleneck (through providing multiple clustered instances of the same component) and (ii) distribution of the components of a single kernel over multiple machines through distributing its components. Figure 2 depicts the logical architecture of a single triplespace kernel.

3.2 Kernel Components

For external communication, a triplespace kernel exposes three interfaces, each geared towards a certain kind of client. Triplespace clients interact with Triple Space through the *Triple Space API* [17] component, which implements the TS API operations. A client invokes a triplespace operation by submitting a request to the TS API implementation, where the TS API operation is translated into an internal request message that is communicated between kernel components. When the requested operation has been processed, the operation's result is returned to the client. In order to integrate triplespace with Web service technologies, a *Web Service API* operating on top of the TS API is provided. It supports (i) publication of Web service descriptions, (ii) discovery of Web services based on their published descriptions and (iii) invocation of Web services over a triplespace based Web service transport binding. Management applications (e.g. for administration purposes) can interact with the *Management API* which is responsible for setting up initial data structures in the integration middleware that are needed for interaction between kernel components and exposes basic management functionality such as initiating kernel startup or shutdown. Persistent storage of data at triplespace kernels is provided through RDF stores which are connected to a kernel by the *Triple Store Adapter* component. The triple store adapter (i) abstracts from different RDF stores and their APIs and (ii) enables transparent distribution of data across a number of physical RDF stores. Apart from hosting the triplespace data, kernel components can use the RDF store through the triple store adapter to persist configuration or runtime data. Security policies are enforced by the *Security Manager* that ensures that all data exchanges across kernel boundaries adhere to specified security policies. The *Mediation Manager* provides functionality for semantic mediation of incoming and outgoing data though execution of either internal transformation logic (e.g. evaluation of transformation rules) or execution of external mediation services. The *Metadata Manager* manages and provides access to a kernel's knowledge about itself (e.g. subspaces, access statistics to triples) and about the global triplespace infrastructure (e.g. other kernels, clustering and routing information). It also provides reasoning capabilities to infer new information from the triplespace ontology that describes a kernel in a semantic description format. The *Query Processor* is responsible for decomposing a query to parts that are satisfiable by the local data store and to parts that must be forwarded to other kernels in order to fulfill the query in its entirety. In order to achieve this, the query processor closely interacts with the distribution manager and the triple store adapter In order to support transactional processing of operations, the *Transaction Manager* implements both local transactions and acts as a transaction coordinator for distributed transactions between multiple kernels. The *Distribution Manager* connects a single kernel to the global space infrastructure. It implements lookup functionality to find other kernels based on triplespace ontology instance data and carries out communication with these kernels. It also takes care of request routing and retrieval of distributed, semantically clustered data. Three distribution mechanisms have been identified and are discussed in Section 3.4.

3.3 Triple Space Kernel Prototype

Currently there is a first prototype implementation of a triplespace kernel available[4]. The latest version of the prototype supports all specified coordination operations and allows associative retrieval based on single triple pattern templates and on SPARQL [16] based templates. The tuplespace functionality is provided

[4] TripCom, http://sourceforge.net/projects/tripcom

by a tuplespace middleware system which serves as a communication bus for the internal triplespace components. For SPARQL based query on the RDF data, the prototype is integrated with the ORDI framework[5]. ORDI is an abstraction layer for heterogeneous ontology models, reasoners, and data storages. For the prototype, it is deployed using the default semantic repository OWLIM[6] which acts as a storage and inference layer for the Sesame RDF store[7]. Employment of ORDI and OWLIM provides a high performance and scalable storage and query solution for processing RDF data. The current prototype can be used to realize distributed applications based on a single triplespace kernel, shared between multiple clients. However, the prototype is in process of being implemented and further tests and evaluations are still to be carried out.

3.4 Triple Space Distributed Architecture

In order to achieve the desired globality in a scalable manner, it is necessary to distribute the data and information sources, i.e. the spaces, over multiple physical machines, i.e. multiple kernels. At the same time respective discovery procedures must be defined, as it is very important that distribution algorithms match the corresponding discovery approaches. Figure 3 describes the relationships between the main entities in a distributed triplespace setting. The central resource are obviously the spaces, where the data, in form of RDF triples, resides. Spaces provide a first step towards a scalable infrastructure by naturally gathering related data and users. Spaces establish a network of virtual information sources that are targeted at a particular topic or a particular user group and hence guarantee at least local scalability, and completeness with respect to the closed world of a space; e.g. a number of business entities decide to collaborate via a triplespace and create their own cooperation platform. In order to provide more expressivity to the set of virtual spaces, any triplespace can have multiple subspaces that are then organized in a tree structure.

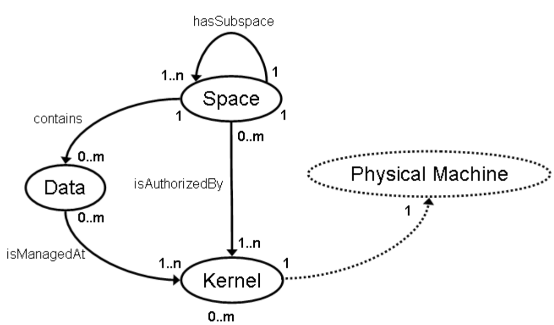


Fig. 3. Triple Space Meta Model

[5] ORDI, http://ordi.sourceforge.net/
[6] OWLIM, http://www.ontotext.com/owlim/
[7] Sesame RDF Store, http://www.openrdf.org

From a client perspective all operations are executed on a space, which in turn must be addressable; therefore triplespaces are assigned to unique identifiers in form of a URI. Still, the actual implementation of the space is provided by the TS kernel, which is formed by the components described in Section 3.1. A kernel is thus the set of software components that provide the functionality of a space and is deployed on a physical machine which of course is also attached to a network address (e.g. an IP address). In consequence, clients that communicate with a space, do not necessarily know where in the network the data resides, nor actually which kernel provides the perceived functionality. Moreover, to pay tribute to the assumed openness of Triple Space installations, and in order to deliver a means for distributedness, a space can be hosted, or authorized, by multiple kernels, while at the same time a single kernel can host multiple spaces.

Consequently, the semantic data is distributed amongst co-authoring kernels. This provides furthermore an important tool for load balancing which is an effective measures for better performance, availability and also reliability. As a result, clients connecting to their triplespace kernel need to be able to discover other kernels in the network on which an operation can eventually be executed, as it hosts the relevant space and thus the targeted data. There are two ways of resolving the target kernel, which are distinguished by how much information a client provides about the information source, i.e. the space of interest.

In case a client knows the identifier of the space it wants to interact with, the space identifier needs to be resolved to the address of the kernel that can interpret the request and return the demanded data. This approach is very similar to the domain name resolution procedure of the World Wide Web, and hence we apply DNS-like techniques too. The advantage of this approach is the inherited scalability of DNS, and the application of the local scalability mentioned above. However, it restricts the user in that it is necessary to know the target space. Moreover, it is a priori not possible to profit from the immense knowledge that resides in the Triple Space, as only the target space is considered. Enhancing these procedures with means like semantic routing (see next paragraph) would diminish this drawback.

The second approach is necessary in case that the client does not or cannot specify the identifier of the desired target space. In this situation the operations must be performed against the content of the virtual global triplespace. Obviously, it is not possible to query the entire Semantic Web, and hence it will never be possible to query the entire Triple Space infrastructure. Still, the current work proposes means to resolve such request with a best effort discovery algorithm. A combination of DHT-like indexes and local semantic routing tables allows any TS kernel to discover relevant kernels and spaces to forward user request too. This procedure ensures the delivery of a semantically correct answer - not necessarily precisely the requested triples, but at least triples reflecting the requested knowledge. This approach does in consequence by no means guarantee completeness compared to the immense knowledge stored in the virtually global space, but has the advantage that users can browse the Triple Space without limiting themselves to a particular predefined subset of the knowledge.

4 Related Work

Based on the initial ideas for combining tuplespaces with the semantic web presented in [4], several systems have been designed and to some extend been implemented: within the project TSC, a tuplespace based communication middleware for Semantic Web services has been developed [5]. Semantic Web Spaces extend the Linda coordination model for exchanging semantic data on the Web [18]. sTuples extend JavaSpaces in order to support semantic data in tuples for use in pervasive environments [12]. CSpaces aim at applications apart from Semantic Web services, e.g. distributed knowledge management [13]. A detailed and comparative analysis of these systems is given in [14], yet none of them provide configurable QoS – a key feature of our approach which is characterized by the notion of sub-spaces that are able to provide QoS as known from traditional EAI middleware, and in addition allows for assigning distinct sets of QoS requirements, so called configurations, to individual sub-spaces. [1] presents a minimal architecture for triplespace computing, based on the principles of the Web and hence with strong focus on scalability. However, it is limited to searching data on single kernels. In order to search the entire network, a global index is required (e.g. the equivalent to a Web search engine).

Several researchers have exploited the P2P approach for distributed tuple spaces. [15] examines the use of coordination-based middleware in a heterogeneous P2P environment and investigates an architecture for decoupling the client-server model within the real world application domain of telecoms management. A tuple space on top of P2P systems for implementing the coordination of peers in fully distributed network games is described in [19]. MobiSpace is a system targeted at mobile distributed environment and implemented in Java 2 Micro Edition [6]. Local replicas are updated via a central server over GPRS or across mobile units using Bluetooth. DTuplesHPC is a Linda like P2P tuple space middleware for use in high performance computing, built on top of distributed hash tables [10].

5 Conclusion

In this work, the concept of Triple Space – a semantically enhanced, distributed tuplespace with configurable QoS – was introduced. The need for providing different QoS within a single middleware system was motivated by two real world scenarios that require disparate and even mutually exclusive QoS for different use-cases but share common data. The solution as presented in the paper is based on the aforementioned Triple Space that provides users with shared data access based on pre-defined sets of QoS, called configurations. Our model foresees for instance sub-spaces as *islands of strong QoS* for interactions that require ACID transactions, and provides an umbrella space for all sub-spaces that allows large-scale access to the data from individual sub-spaces. Subsequently, the architecture of a single Triple Space kernel, along with a description of supported configurations was presented, followed by a description how these kernels interact in order to build a distributed system.

References

1. Bussler, C.: A Minimal Triple Space Computing Architecture. In: 2nd WSMO Implementation Workshop (June, 2005)
2. de Francisco, D., et al.: Using triple spaces to implement a marketplace pattern. In: Procceedings of the 1st SBC Workshop in 1st European Semantic Technology Conference (ESTC 2007) (May 2007)
3. Fensel, D.: Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. Springer, Heidelberg (2001)
4. Fensel, D.: Triple-Space Computing: Semantic Web Services Based on Persistent Publication of Information. In: Aagesen, F.A., Anutariya, C., Wuwongse, V. (eds.) INTELLCOMM 2004. LNCS, vol. 3283, pp. 43–53. Springer, Heidelberg (2004)
5. Fensel, D., Krummenacher, R., Shafiq, O.: TSC - Triple Space Computing. e&i Elektrotechnik und Informationstechnik (forthcoming) 124(1/2) (2007)
6. Fongen, A., Taylor, S.J.E.: MobiSpace: A Distributed Tuplespace for J2me Environments. In: Parallel and Distributed Computing and Systems PDCS 2005 (Nov, 2005)
7. de Francisco, D., et al.: Towards a digital content services design based on triple space. 10th Int'l Conference on Business Information Systems (BIS) (2007)
8. Gelernter, D.: Generative communication in Linda. ACM Transactions on Programming Languages and Systems 7(1), 80–112 (1985)
9. Hohpe, G., Woolf, B.: Enterprise Integration Patterns. Designing, Building and Deploying Messaging. Addison-Wesley, Reading (2003)
10. Jiang, Y., Xue, G., Li, M., You, J.: DTuplesHPC: Distributed Tuple Space for Desktop High Performance Computing. In: Jesshope, C., Egan, C. (eds.) ACSAC 2006. LNCS, vol. 4186, pp. 394–400. Springer, Heidelberg (2006)
11. Kalfoglou, Y., Schorlemmer, M.: Ontology mapping: The state of the art. Knowledge Engineering Review 18(1), 1–31 (2003)
12. Khushraj, D., Lassila, O., Finin, T.W.: sTuples: Semantic Tuple Spaces. In: Proc. 1st Ann. Int'l Conf. on Mobile and Ubiquitous Systems: Networking and Services, pp. 268–277. IEEE Computer Society Press, Los Alamitos (2004)
13. Martín-Recuerda, F.: Application Integration Using Conceptual Spaces (CSpaces). In: Proc. 1st Asian Semantic Web Conf., pp. 300–306. Springer, Heidelberg (2006)
14. Nixon, L., Simperl, E., Krummenacher, R., Martin-Recuerda, F.: Tuplespace-based computing for the semantic web: A survey of the state of the art. Knowledge Engineering Review (2007)
15. Parker, D., Cleary, D.: Building Richer JXTA Applications with Collaborative Spaces in a Peer-2-Peer Environment. In: Proceedings of the 38th Annual Hawaii International Conference on System Sciences, 2005 HICSS 2005 (Jan, 2005)
16. Purd'hommeaux, E., Seaborne, A.: SPARQL Query Language for RDF (2006)
17. Simperl, E., Krummenacher, R., Nixon, L.: A Coordination Model for Triplespace Computing. In: 9th Int'l Conference on Coordination Models and Languages (2007)
18. Tolksdorf, R., Nixon, L., Bontas, E.P., Nguyen, D.M., Liebsch, F.: Enabling real world Semantic Web applications through a coordination middleware. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532, pp. 679–693. Springer, Heidelberg (2005)
19. Wang, C.-W., Hsiao, H.-C., Sun, W.-H., King, C.-T., Sun, M.-T.: Building a tuple space on structured peer-to-peer networks. The Journal of Supercomputing 40(3), 213–227 (2007)

Performance and Architecture Modeling of Interoperability System for SME's

Mindaugas Kiauleikis[1], Valentinas Kiauleikis[1], Claudia Guglielmina[2], Karsten Tolle[3], and Nerijus Morkevicius[4]

[1] Kaunas University of Technology, Computer Engineering Department, Studentu st. 50, LT-58631 Kaunas, Lithuania
[2] TXT e-solutions, Via Frigia, 27, 20126 Milano, Italy
[3] University Frankfurt am Main - DBIS, Robert-Mayer-Str. 11-15, 60325, Frankfurt am Main, Germany
[4] Kaunas University of Technology, Department of Applied Mathematics, Studentu st. 50, Kaunas, Lithuania

Abstract. SMEs' needs to have sufficient information system functionality for affordable price were the main reason for development of business integration. Enterprise Service Bus (ESB) development and technology creation helped to better fulfill modern needs of SMEs. One of ESB variants – Application Bus for InteroperabiLITy In enlarged Europe SMEs (ABILITIES) – is described in this paper. Problems of creation, architecture and performance issues of such interoperability infrastructure are discussed. Colored Petri Net model for performance evaluation of such infrastructure is presented and evaluated.

Keywords: Interoperability, eBusiness, UBL, SMEs Networks, ESB, ABILITIES.

1 Introduction

Enterprise interoperability is a key factor in enabling enterprises to harness the full potential of IT and related services. Investment in R&D and innovation of SMEs interoperability is one of the key pillars of the i2010 initiative [7] that was launched in June 2005. European Commission funded a number of research issues (ATHENA, TrustCoM, Interop, FUSION, ABILITIES and other) in enterprise interoperability for modeling interoperability infrastructure, networked enterprises, and creating interoperability tools using architectural guidelines that support the interoperability of inter enterprise systems (RosettaNet Framework, the OAG Integration Specification). Among the most well known is ebXML as well as a set of specifications supported by OASIS (UBL, WS-BPEL) [3, 5] and proposed by a large group of business, government standards committees and academics.

Looking from Lithuanian business case (there is evidence to suggest that in other new EU member states too) global data transfer network infrastructure is well-developed and allows businessmen to go out of country borders freely making and keeping e-business contacts. But initiative of businessmen is restricted by four evident barriers:

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 345–356, 2008.

1. **Language.** All collaboration partners in different countries use their own languages so business relations are encumbered by translations, explanation of meanings and of course misunderstanding.
2. **Law.** Each country has its own law according which businessmen register and run their activities. Law describes processes and documents of interoperability between businessmen and state institutions. Each must be adjusted when business goes out of country's borders.
3. **Document forms and mandatory data.** In Lithuania, there are regulated forms of documents that are in strict accountability in state institutions. Other countries usually have their own strict accountability documents with different form and data. Inter-organizational accounting documents usually have a free form, but some large organizations, which deal with many small ones, usually create their own standards and require everyone to comply with them. Document forms and data must be reconciliated if documents are passed from one country to another or even from one company to another.
4. **Business process.** Same as document forms business process in each Lithuanian business enterprises is specific. So interoperability system must be flexible not to conflict with business processes of different enterprises inside or outside of the country.

Each barrier prevents e-business growth not only in international scale but also between Lithuanian businessmen. The aim of this research is: 1. to define which additional internet services are needed to create interoperability infrastructure; 2. to create a model of such services, examine computational resource needs for interoperability infrastructure and present methodology for modeling of architectural implementation impact on overall performance of system.

This research is based on architectural and technical concepts of EU funded project ABILITIES (Application Bus for InteroperabiLITy In enlarged Europe SMEs). The short description of main ideas of ABILITIES is described in the section 2. ABILITIES uses the XML based Universal Business Language (UBL) [3, 5] described in section 3 of this paper.

2 ABILITIES Project

The EU funded project ABILITIES addresses SME interoperability technologies and solutions. Interoperability identifies the ability of new generation Enterprise Applications (EA) to simplify, speed-up and rationalize the implementation of EAI projects involving them. The analysis of research about enterprise interoperability led to the identification of the two main research priorities: adaptation of UBL documents [3] to the requirements of SMEs in Enlarged Europe; and federated architecture for interoperability with intelligent, adaptive business documents; reconciliation and integration of state-of-the-art languages and standards for Business Process Management and Service Orchestration with new research results in the field of Business Documents and Messages.

The solution proposed for supporting both interoperability levels is a blended architecture which could join the peculiar advantages of message-based Service Oriented Architectures in the business documents format context and intelligent systems for mutual understanding and agreement on document contents. The overall goal is not to change the business, processes or existing systems, but to adopt them.

The heart of the ABILITIES Architecture is the ABILITIES Interoperability Bus (AIB), based on an open source Enterprise Service Bus. All core components that implement the interoperability support for the collaborating enterprises are connected to this ESB, as shown in Fig. 1. The architecture integrates and supplies a number of functions.

- **Collaboration Configuration** – the communication between SMEs purchase order details and defining specifications of the goods or services is essential. With this configuration it can be defined in what cases a collaboration session between two companies should be started and what channel should be used, to handle open issues.
- **Negotiation Rules Handling** – each company can define common rules it follows and even special rules for certain partner companies. These rules will be used to avoid or at least to shorten bargain by given rules from each company.
- **Process Design** – processes of companies differ depending on size, domain, location, etc. Interoperability system will have a defined Global Process designed, capturing the peculiarities of the location and domain. Each company therefore needs to define Private Processes as sub blocks of the global one.
- **Roles Management** – the employees of the company need to be adjusted to roles and/or tasks defined during the process design, to enable the system to transmit it via the appropriate channel to the correct person.
- **Business Documents Definition and Adjustment** – each SME relies on its specificities, adherence to local or national regulations and norms, as well as internal norms. Starting from a common UBL base of Business Documents, ABILITIES supports derivations and adjustment of the standard UBL messages – called UBLtc. This approach ensures domain or regional or custom localizations while at the same time ensuring interoperation at the level of business documents.

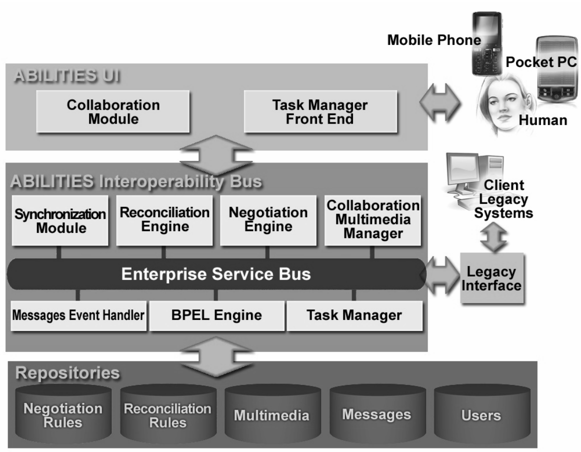


Fig. 1. Internal view of the ABILITIES Interoperability Bus (AIB) showing the core components

The core functionality of the system is to exchange the messages between the two trading parties. While passing the message from one party to another the message will be transformed and might be modified based on the defined negotiation rules. The transformation follows the unified approach translating the sender message according to the ontology adjustments performed by the sending and receiving party (reconciliation engine). Additionally, the federated approach is used to negotiate the message business content (price, quantity, warranty duration, delivery date, quality). This is the task of the negotiation engine.

3 UBL and ABILITIES Message Structure

Whenever people trade, there is the need to exchange information. Most times this is done by business documents (BODs). These BODs contain contractual information that also needs to fit to current legislation issues. This was already the case thousands of years ago, when people exchanged stone tablets containing the order or invoice information. Nowadays electronic BODs are used that can be processed and manipulated automatically, as we do within the ABILITIES solution.

When we talk about interoperability the understanding of the content of the BODs is needed. Not only by humans, but also by automated systems. Therefore, the used encoding format – defining what content is placed into the document and how it can be extracted – of the BODs plays an important role. There exist different encoding formats for electronic BODs. The most common one, especially in Europe, is UN/EDIFACT by UN/CEFACT. However, EDIFACT is a very monolithic, inflexible and complicated standard, which results in very high introductory and maintain costs. For SMEs the usage of EDIFACT is therefore a no-go, especially when they have many and frequently changing trading partners.

In ABILITIES we therefore selected the Universal Business Language (UBL) as our basis for document encoding using latest technologies. The Universal Business Language (UBL) was released as version 1.0 by OASIS in 2004 as a further development of the Common Business Language (CBL). In December 2006 version 2.0 was approved as a major release and is the current actual version. UBL 2.0 defines business documents for the business areas: Sourcing, Ordering, Invoice and Fulfillment. These documents are build-up out of a library of XML Schemas for reusable data components. This way UBL is the first specification that implements the Core Components Technical Specification (CCTS) of ebXML. Alternatively RosettaNet provides document encoding based on older XML/DTD technology.

To sum up, the main reasons for selecting UBL were:

1. Simplicity – the design objective of UBL was to follow an 80/20 rule to provide as much components and documents as needed with a very low level of complexity [6].
2. Customizability – UBL is generic. It therefore provides the possibility to confine, extend or even to create your own documents.
3. Based on XML – UBL is based on XML and XML Schema. Therefore, it can be created, processed and validated easily with available and cheap standard tools. Additionally an increasing number of experts for XML are available.

In order to be able to overcome the mentioned barriers for SMEs – as described in section 1 – these features are needed. Especially being able to customize your document due to legislation or domain needs is extremely important. In AIB we have with the Reconciliation Engine a separate module trying to map different customized documents in order to minimize the work needed to interconnect two companies and therefore to increase interoperability.

From the side of OASIS hosting the UBL standard, they are trying to constitute subcommittees for different countries, called Localization Subcommittees (LSCs). These LSCs are in charge of translating the different elements of the common library and the documents, as well as applying peculiarities of the country. This reduces the effort of single SMEs to hook to UBL and helps them to overcome the mentioned barriers 1) Language, 2) Law and 3) Forms of documents.

Last but not least we can state that the usage of UBL is future-oriented. There had been a memorandum of understanding between OASIS and CEFACT dated April 2006, where UBL is recognized as starting point for future developments of EDIFACT [5]. This ensures some stability and reduces the risk of SMEs to put money on a standard that might disappear after a short period.

4 Model of Interoperability System

An abstract model of Interoperability System (IS) based on "real world" ABILITIES system is described in this section. Main components of the model are similar to AIB.

All AIB modules can be divided into two classes: design time modules and run time modules. Design time modules are: Collaboration Configuration Manager, Process Designer, Ontology Delta Editor, Negotiation Designer, Multimedia Content Manager, and User Manager. Negotiation Rules, Reconciliation Rules, Multimedia and Users repositories are used to store required data for normal work of these modules. ABILITIES end users use these modules to prepare data and rules for normal ABILITIES system work at runtime phase.

The runtime modules of ABILITIES system receive, handle, convert, process (according to data prepared at design time and stored in corresponding repositories) and send messages to destination points. The runtime modules include: Message Event Handler, Synchronization module, Reconciliation Engine, Collaboration module. Repositories with necessary data are used at run time: Negotiation Rules, Reconciliation Rules and Messages repositories.

Performance of IS at design time is not the option for optimization because it requires end user interaction. On the other hand, performance of runtime modules is important for purposes of design of "real world" implementations of such system.

Proposed IS model has only runtime modules. Collaboration module (CM) needs end users' interactions so it is assumed that message leaves IS when it reaches CM. The runtime modules use some repositories, only two of them are included – Reconciliation and Negotiation rules repositories. Other repositories are either used at design time (User repository) or are accessed directly by end users (Multimedia repository), so they do not add direct impact on time the message spends in interoperability system. Message flow through such system is shown in Fig. 2.

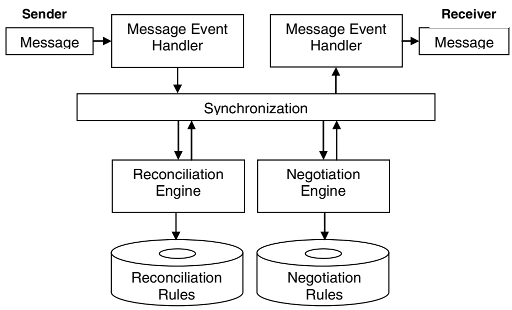


Fig. 2. Model and message flow in Interoperability System

Message Event Handler waits for message entering the IS. Then it handles the message and delivers it to Synchronization module. Synchronization module decides which message transformations are needed and sends it to the Reconciliation and/or Negotiation engines (if needed). The Reconciliation engine performs semantic reconciliation, finds and loads semantic reconciliation rules, translates UBL business documents, etc. Negotiation engine applies stored negotiation rules and transforms the message. Finally, Synchronization Engine delivers message to Message Event Handler, which converts message into the appropriate format and sends it to destination party.

Petri net model presented in the next section simulates the behavior of IS at run time from the moment the message enters the IS until it is processed and sent to the receiver.

5 Petri Net Model of Interoperability System

Abstract model of interoperability system could be used to create colored Petri net [1], [2] model of such system. Basic scheme of one variation of such model is shown in Fig. 3.

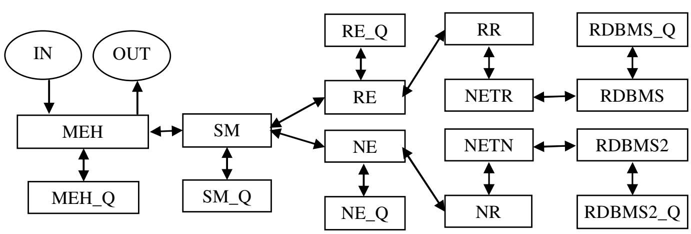


Fig. 3. The basic scheme of Petri net model of interoperability system (Second model)

The model consists of four main modules: Message Event Handler (MEH), Synchronization module (SM), Reconciliation engine (RE) and Negotiation engine (NE). Messages enter system (IN place) and are processed in these modules as needed. Each of these modules has its own queue (MEH_Q, SM_Q, RE_Q and NE_Q, respectively). Negotiation and Reconciliation engines use data from two repositories: Reconciliation repository (RR) and Negotiation repository (NE). This version of IS model assumes that both repositories are stored in different remote relational databases. Network access delay is modeled by two modules (NETR and NETN). Repositories' data are physically stored on two different databases (RDMS and RDMS2) with queues RDBMS_Q and RDBMS2_Q respectively.

IS model has several modeling parameters. Parameter IncInt with the mean IncIntm and exponential distribution represents time interval between two message arrivals. Each message has two additional parameters (r and n) which are binary numbers and may have values "1" or "0". If message's parameter r has value "1" then this message must be processed in Reconciliation engine. Parameter n turns to "1" when message must be processed in Negotiation engine. Values of these parameters are generated randomly, are independent and have probabilities to be "1" Rp and Np, respectively. All message handling delays have Poisson distributions with means Mehd, Smd, Red, Ned, Rrd, Nrd, Netrd, Netnd, Rdbmsd and Rdbms2d for MEH, SM, RE, NE, RR, NR, NETR, NETN, RDBMS and RDBMS2 respectively. Processed message is sent to receiver (OUT place).

Color Petri net model was created using CPN tools software. Each module in Fig. 3 is modeled by one or several Petri net pages, connected using fusion places and hierarchical structure. Each token in Petri net represents one message (document) in interoperability system.

Message event handler Petri net page is shown in Fig. 4.

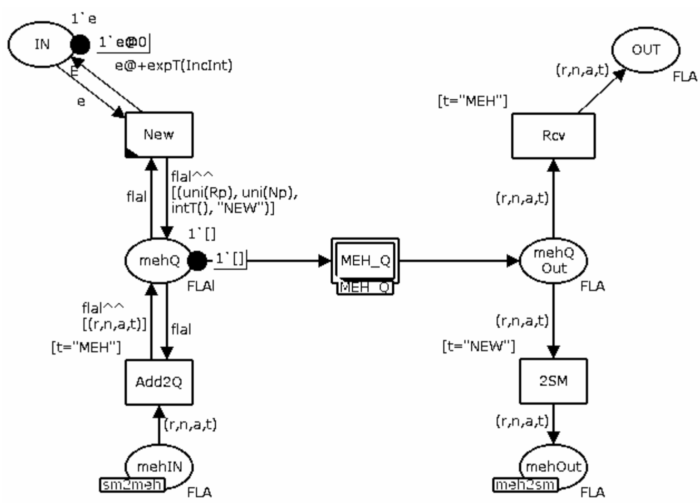


Fig. 4. Message event handler Petri net page

Message Event Handler Petri net page is defined by a tuple
$N = (P, T, Pre, Post, C, cd)$,

where

- P is a finite set of places,
- T is a finite set of transitions,
- C is set of color classes,
- $cd : P \cup T \to C$ is the color domain mapping,
- $Pre, Post \in B^{|P| \times |T|}$ are backward and forward incidence matrices.

Set of places of MEH Petri net:

$$P = (IN, OUT, mehIN, mehOUT, mehQ, mehQOut) ,$$

where *IN*, *OUT* – entering and exit points of Abilities system, *mehIN*, *mehOUT* - entering and exit points of MEH to other modules of IS and *mehQ*, *mehQOut* – entering and exit points to query (MEH_Q) of Message Event Handler.
Set of transitions of MEH Petri net:

$$T = (New, Add2Q, MEH\_Q, Rcv, 2SM) ,$$

where *New* is transition which puts new arrived documents to *MEH_Q, Add2Q* – puts processed documents to *MEH_Q*, *Rcv* – sends processed documents to receiver, *2SM* – sends new arrived documents to Synchronization module, MEH_Q – represents Petri net of queue of Message Event Handler.
Color sets of MEH Petri net:

- *colset REF=int* – Reconciliation engine flag,
- *colset NEF=int* – Negotiation engine flag,
- *colset AT=int* – Arrival time,
- *colset TIP=int* – Message type,
- *colset FLA=product REF*NEF*AT*TIP timed* – Message,
- *colset FLAl=list FLA* – list of messages,
- *colset E=with e timed* – Empty (used for generation of messages).

$cd : P \cup T \to C$ color domain mappings are not defined, because of big number of tokens used in simulation.

Other parts of IS model are not presented here. All queues used in model are FIFO queues with unlimited length and monitoring facilities. The created Petri net allows monitoring of maximal and average queue lengths for all modules. Also average, minimal and maximal message processing time may be measured.

6 Results of Modeling

Modeling objective was to find out the performance of IS at run time under the different load of messages. It was also interesting to find out what influence (if any) the change of interoperability system's architecture could provide. For this purpose three versions of IS model with slightly different architectures were created. First one assumes that both repositories are located in the same computer as the IS software (Fig. 5).

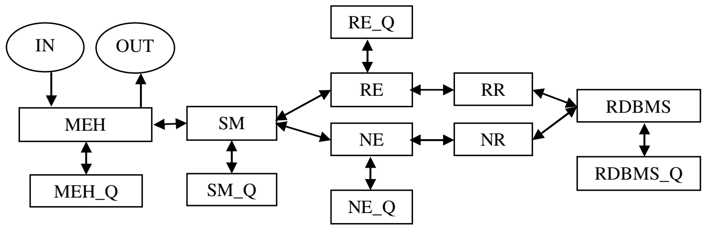


Fig. 5. IS model with all repositories stored on the same computer as IS (First model)

In such case there are no delays caused by the use of computer network, but usage of one database causes IS bottleneck in repositories. More detailed description and performance testing results of this model can be found in [4].

One of possible solutions to avoid bottleneck in repositories is to use two different databases running on two different computers. In such case we get model of IS shown in Fig. 3 (second model). Initial performance modeling showed that additional delays in network subsystem cause that overall system performance (average message processing time) increase is very little.

To achieve additional performance gain, IS architecture may be changed to one presented in Fig. 6 (third model).

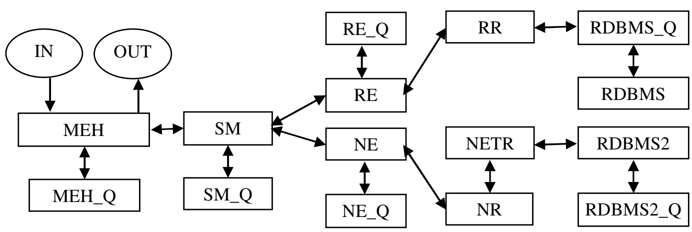


Fig. 6. IS model with one repository stored on remote computer (Third model)

Now Reconciliation repository is stored on the same computer as main IS. This repository is accessed (according to our testing parameters [4]) more frequently, so we can save more time on network delays. On the other hand, Negotiation repository is stored on remote computer, but it should cause less overall performance drop as negotiation is needed more rarely.

Test modeling was performed using parameters presented in Table 1. All times are presented in relative time units (for example milliseconds).

Table 1. Parameters for test modeling

Rp	Np	Mehd	Smd	Red	Ned	Rrd	Nrd	Rdbmsd	Rdbms2d	Netrd	Netnd
0.9	0.3	30	50	200	200	50	50	500	500	100	100

Dependences of the average message processing time on inter-message arrival time using three different IS architectures are presented in Fig. 7.

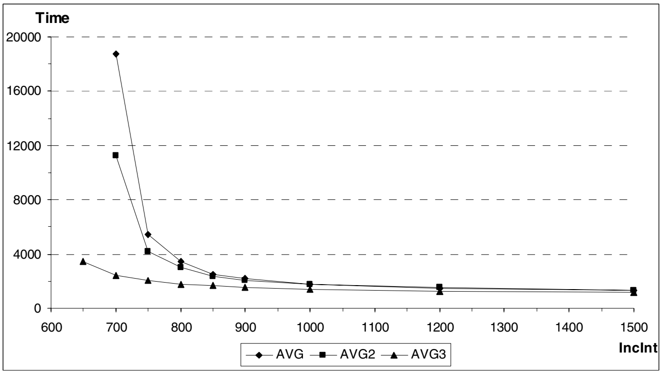


Fig. 7. Dependences of message processing time on inter-message arrival time in three different models

Here AVG, AVG2 and AVG3 are average message processing times using first, second and third model respectively. One can clearly see than under big load of messages third architecture is considerably more efficient.

Dependences of maximal count of messages in Reconciliation repositories on inter-message arrival time are presented in Fig. 8. Where ReMax1, ReMax2 and ReMax3 are the maximal lengths of the queues in Reconciliation engines, using first, second and third IS model respectively.

Finally we present more detailed test modeling results of IS using one remote database (third model). These are the best ones (performance vise) of three systems.

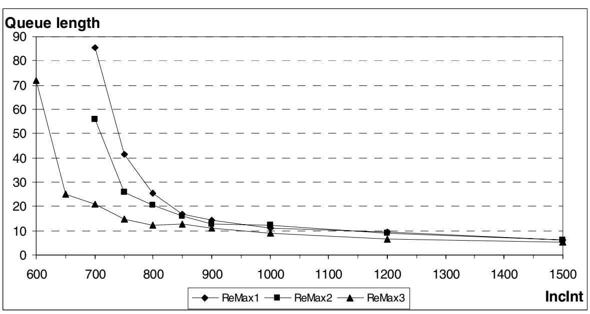


Fig. 8. Dependences of maximal count of messages in reconciliation engine queues on inter-message arrival time (IncInt) in three different models of IS

Fig. 9 presents dependences of average message processing time on inter-message arrival time. AVG is an average message processing time in system. T00, T10, T01 and T11 are average processing times of four different message types: T00 represents messages which are not handled at all, they are simply sent to receiver; T01 represents messages which are handled only in Negotiation engine; T10 are handled only in Reconciliation engine, etc.

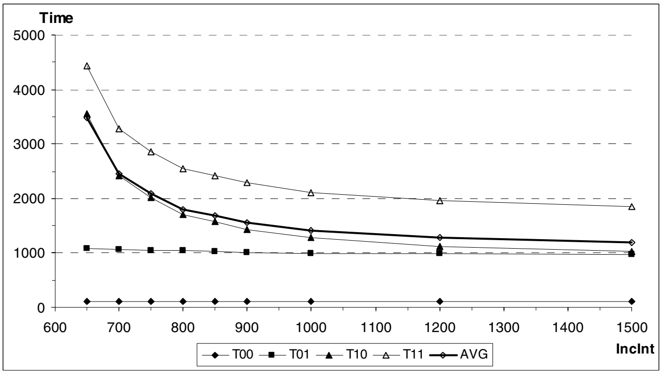


Fig. 9. Dependences of message processing time on inter-message arrival time (Third model)

Dependences of maximal count of messages in repositories queues on inter-message arrival time (IncInt) are presented in Fig. 10. Here ReMax is the maximal length of the queue in Reconciliation engine, NeMax – the maximal length of the queue in Negotiation engine.

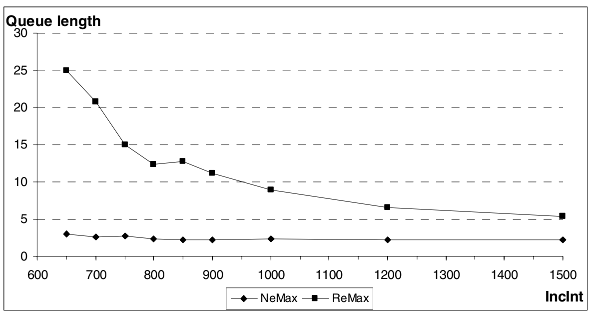


Fig. 10. Dependences of maximal count of messages in queues on inter-message arrival time (Third model)

7 Conclusions

1. ABILITIES project is perspective for creation of interoperability infrastructure. It should solve businessmen interoperability problems by introducing needed functionality by extending web services without big additional expenses for purchase and maintenance of software. To cheapen, but at the same time make more effective IT implementations is the main advantage of ABILITIES.
2. Created Petri net model can be used to answer questions on IS performance, bottlenecks and scalability. It can help to estimate resources needed for wider region interoperability.

3. The created model can be used to find out which IS modules cause biggest delays in whole message flow. According to this information IS architecture can be changed to suit new expectations.
4. Real world ABILITIES system technical parameters are unavailable at this time, so testing of the created model can't be finished until all time characteristics of message handling in various modules are available.

References

1. Jensen, K.: An Introduction to the Practical Use of Coloured Petri Nets. In: Reisig, W., Rozenberg, G. (eds.) APN 1998. LNCS, vol. 1492, pp. 237–292. Springer, Heidelberg (1998)
2. Girault, C., Valk, R.: Petri Nets for Systems Engineering. Springer, Berlin (2003)
3. Gertner, M., Gutentag, E., Gregory, A.: Guidelines For The Customization of UBL v1.0 Schemas. In: OASIS Working Draft (2004)
4. Guglielmina, C., Janavičiūtė, A., Kiauleikis, M., Kiauleikis, V., Morkevicius, N.: Performance modeling of interoperability system for SME's. Information technology and control 35(4), 424–430 (2006)
5. McGrath, T.: UBL and UN/CEFACT - adding sanction to traction. In: Presentation at UBL International 2006, 13-17 Nov. Copenhagen, Denmark http://www.ublconference.com/200611/programme.html#ublcefact
6. Bosak, J., McGrath, T., Holman, G.K.: Universal Business Language v2.0. OASIS Standard 12 December 2006, http://docs.oasis-open.org/ubl/os-UBL-2.0/UBL-2.0.html
7. i2010 - A European Information Society for growth and employment, http://ec.europa.eu/information_society/eeurope/i2010/index_en.htm

Measurement of Business Process Orientation in Transitional Organizations: An Empirical Study

Uma Kumar, Kayvan M. Lavassani, Vinod Kumar, and Bahar Movahedi

Sprott School Of Business, Carleton University,
1125 Colonel By Drive, Ottawa, Canada

Abstract. With regards to the benefits associated with the business process orientation, only a small body of research in this field has specifically addressed the issue of the measurement of process-orientation in organizations. To help fill this gap, the primary focus of this study is to explore the measurement of business process orientation in transitional organizations by providing a framework for measuring process orientation, from managerial perspective, and applying this measurement in an empirical study. Based on the transition framework of this study the transition process from functional organizational structure to a process-oriented organizational structure starts with the implementation of the ERP system by the organization. The results of the study show the positive relationship that exists between the level of process-orientation in the organization and the success of ERP systems implementation and utilization.

Keywords: Business Process Orientation, Enterprise Resource Planning (ERP) systems, Measurement System, ERP Implementation.

1 Introduction

During the early 1990s the concept of business process-orientation attracted the interest of many researchers, and practitioners. Many companies started to realize the benefits associated with implementing process-oriented structure. For example, IBM conducted a review of its business processes and realized that majority of its cooperate customers were increasingly operating on a global basis. Therefore, IBM started to standardize its business processes, and initiated a common business process approach for all its day to day activities. The benefits have astonishing, with a 75% reduction in the average time to market for new products, a rapid increase in on time deliveries and customer satisfaction, and cost savings of over 9 billion dollars. There is increasing evidence from this and many other successful companies that provide evidence for exceptional competitive advantage results when a firm combines its organization's assets and skills (e.g., marketing capabilities, innovation) with process-orientation (Vathanophas, 2007; Regev *et al.*, 2005; Sara *et al.*, 2004; McCormack and Johnson, 2002; and Llewellyn and Armistead, 2000).

One of the common themes among the available literature about business process-orientation is the notion of process culture with the strong emphasize on customer

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 357–368, 2008.

satisfaction, and outcomes since commitment to process development directly benefits the customers (Balasubramanian and Gupta, 2005; Balzarova *et al.*, 2004). Moreover, a culture of team-orientation, and empowerment of individuals who assist in developing the process-orientation, and value the costumers are the two core concepts of business process-orientation (Biazzo, 2002). Business process-orientation can help reduce the tensions which may exist between different functions in an organization, and also help increase the connectedness across departments (Hammer, 2003, 1996; McCormack, et al. 2002; Lockamy, *et al.*2004; and Llewellyn and Armistead, 2000). All of which result in greater organizational performance both in short and long term. Research also shows implementing business process-orientation can help increase the internal coordination and create stronger team spirit in most companies (Love *et al.*, 2000; and McQueen, 1999). Moreover, process-orientation can dramatically influence the performance of an organization by increasing the employee's feelings of devotion to a common cause and enthusiasm (Regev *et al.* 2005; McCormack, et al. 2002). Furthermore, when jobs become process-oriented in an organization it results in process-oriented authority. This type of authority would persuade employees from different functions to work together toward common goals of organization. Therefore implementing a process-oriented structure in an organization does not require a charismatic leader; instead it can be initiated and developed through organizational dynamics (Shari and Seddon, 2007; Weerakkody *et al.*, 2003; and McCormack, et al. 2002).

With regards to the benefits associated with the business process orientation, only a small body of research in this field has specifically addressed the issue of the measurement of process-orientation in organizations. To help fill this gap, the primary focus of this study is to explore the measurement of business process orientation in transitional organizations by providing a framework for measuring process orientation, and applying this measurement in an empirical study. For the purpose of this study the transition process from functional organizational structure to a process-oriented organizational structure starts with implementation of the ERP system by the organization.

2 Measurement of Process-Orientation in Organizations

Measurement of process-orientation is an "underdeveloped activity of strategic importance in today's process-oriented organizations" (Ljungberg, 2002). Measurement has been identified as one of the four key competences (the other three are positioning, integration, and agility) needed for the achievement of "world class logistics", and it is often pointed out as an area in need for further research (Sara *et al.* 2004). The purpose of process-orientation measurement system is to facilitate the transformation through insuring the alignment of process-orientation with organizational goal (Shari and Seddon, 2007; and Ljungberg, 2002). The organization has to set "Key Performance Indicators" for the implementation process (Balzarova, 2004) and when the measurement system shows any misalignment between outcomes and organizational goals, the transformation team will realize the need for reassessment of 'refined business processes' (Vathanophas, 2007). Also, when measurement system shows alignment between outcomes of the process-orientation and organizational goals then the transformation team will be able to progress the transformation process with more confident, assurance and support, in the latter case, process measurement can support and ease the creation of a process-oriented organization by driving and facilitating business development.

Measures of process-orientated in organizations can be classified into two main groups: the actual *process* which could be described as performance drivers or the *resources* it requires (Ljungberg, 2002). When measurement system shows shortcomings in achieving goals at actual process level, the transformation team should focus on enhancing the outcome of the processes which in turn increases the effectiveness of processes. Three types of structural metrics have been identified for measuring the effectiveness of measurement systems: Complexity, Integration and Dynamism to quantify structural characteristics of information age businesses

Table 1. Measures of Process-orientation in the organization

1 Existence of a systematic decision making process which is defined based on business rules.	**11** The processes being documented properly and documentation is accessible to employees to be used as a reference.
2 Reconfiguring the physical workspace to facilitate process flows within the company.	**12** No misalignment between management's vision and corporate culture in our organization.
3 Process participants at all levels of the organization have a systematic and effective way of communicating with each other.	**13** Executive managers provide adequate input for designing and implementing processes.
4 The majority of processes being either interactive or automated.	**14** Top management is relates strategy and goals to the processes.
5 Existence of continuity and integration in the execution of activities within a process.	**15** Processes to be designed toward satisfying customers' expectations.
6 Processes being visible to relevant stakeholders.	**16** Top management, provides process mapping, in which all sub-processes, information flows, inputs, and outputs together are identified, and described.
7 Carrying out the majority of processes not depends on human discretion and human judgment.	**17** Existence of a measurement system that evaluates the effectiveness of processes across the organization.
8 Ability of execution of several or more than one activities simultaneously.	**18** Provision of adequate knowledge, skills and training for all the employees on how to implement a successful process measurement system.
9 Delays not creeping in due to frequent intervention of humans in carrying out the processes.	**19** Top management and key stakeholders organization to be committed to and support the change in processes due to ES implementation.
10 Existence of an effective process in place for identifying and resolving problems within and between processes.	

(i.e., simplicity, integration, self-learning, agility, and robustness) (Regev *et al.*, 2005). Recognizing the type of effectiveness shortcomings is the key to enhancing the level of process-orientation in an organization. In addition, when shortcoming at resource level is identified (e.g. when the processes employ too much resources for performing tasks) the transformation team should focus on enhancing the processes in order to increase the efficiency of processes. Moreover, Shari and Seddon, 2007 recommend that for achieving efficiency in process-oriented organizations the transformation team should employ process configuration or design information such as activity autonomy and activity automation. One of the most comprehensive measurement systems is developed by Balasubramanian and Gupta (2005) in which they introduce a wide rage of measurement tools for measuring different aspect of process-orientation. Based on the review of the literature we developed a measurement system from organizational and managerial perspectives, which consists of 19 items. These measures are presented in Table 1.

3 Transformation from Functional to Process-Oriented Organization

There are several studies available in the literature describing the transfer process of traditional functional organizations into process-oriented organizations. Figure 1

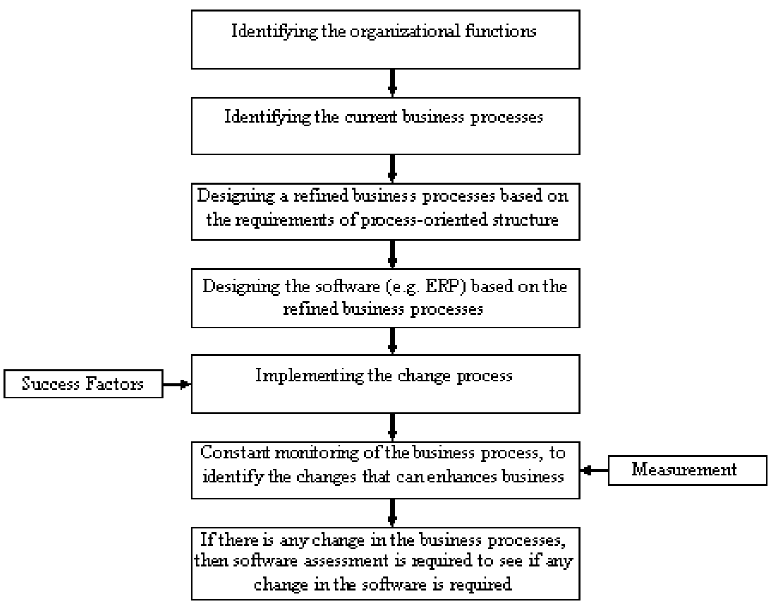


Fig. 1. Transition Model of Functional organization to Process-Oriented Organization. Adopted from Fernandes et al. (2001), modified by the authors.

presents a comprehensive transformation model developed by Fernandes *et al.* (2001). In this model transformation process starts with identifying organizational functions and business processes and ends with monitoring activities, success measurement, and IT infrastructure adoption to the new processes.

In nowadays organizations, the IT infrastructure -namely Enterprise Resource Planning (ERP) systems- is the backbone of organizational operations, and one of the major enabling factors for transforming into a process-oriented organization (Tchokogue *et al.,* 2005). Change implementation and assessment of the change process are the most important parts of transformation of functional organizations into process-oriented ones. In order for the transition to take place successfully the transition team should have an appropriate measurement system in place and also consider the actors which influence success of this transition. As it was illustrated in Figure 1 the transition process involves seven stages. For the purpose of this study we mainly focus on the fifth and sixth stages, namely, “implementing change process”, and “monitoring the business processes”. In the following sections we will explain the role of measurement system in success of organizational transformation to process orientation, and the managerial and operational factors affecting the success of implementation of process-oriented management system in organizations. For assessing the success of ERP implementation we use 14 measures of success

Table 2. Measures of Success of ERP system Implementation

1 Use of project management practices that effectively support the ES implementation	**9** Redesigning processes in conjunction with the ES implementation
2 Training most employees to understand and use end-to-end business processes using the ES	**10** Paying enough by the management attention to restructuring reward and incentive systems subsequent to ES implementation
3 Existence of systems to support, encourage, and reward teamwork and team development	**11**Enough attention to people-based factors that would support the successful use of the deployed processes
4 Existence of uncertainty among the employees about their involvement and role in the change process due to ES implementation	**12** Enough attention has been paid by the management to understand and apply industry best practices for managing the deployment of ES infrastructure
5 Readily shared information for decision making within the company	**13** Providing opportunities for continuous learning about ES functionality in the organization
6 Provision of clear vision and well-defined roles by management in order to eliminate resistance to change	**14** Allocation of sufficient time by the organization for the implementation of ES
7 Enhanced access to organization’s, suppliers’, and customers’ information post deployment of ES	
8 Existence of ambivalence on how organization manages IT investments	

developed based on the literature review and our preliminary interviews. These factors are presented in Table 2. In addition we use the 19 measures of process orientation described in Table 1 for monitoring the business processes.

In this study the success of utilization of ERP was also measures. 13 measures were used to assess the different outcomes of deploying ERP in organization. Table 3 displays the measures of success of utilization of ERP system.

Table 3. Measures of Success of Utilization of ERP system

1 Work has become easier	**9** Access to higher quality data and information about customers and suppliers
2 Employees have more collaboration with other employees in performing their tasks	**10** Access to a more up-to-date and flexible technical infrastructure post ES deployment
3 The jobs are more satisfying for the employees	**11** Major positive changes in the jobs and roles of individuals
4 Ability of the company to better pursue new business opportunities	**12** Employees have more autonomy in decision making that is directly related to their work
5 Clear financial benefits resulting from deployment of ES	**13** The implemented ES helps the company to reach and serve more customers than it previously did
6 Clear cost savings result from deployment of ES	
7 Creation of new sources of revenue resulting from deployment of ES	
8 More flexibility and responsiveness in delivering products and services across the organization	

4 Research Design, Data Collection, and Data Analysis

The transition in the process of moving from functional organizational structure to a process-oriented organizational structure starts with the implementation the ERP system by the organization. The process of implementing the ERP system is a never ending process. However we should note that implementation of ERP system and utilization of ERP system do not happen at the same time. Figure 2 presents the sequence of events in the process of ERP system implementation.

There are three phases during the transition process, namely, before ERP system implementation, during ERP systems implementation, and after utilizing the ERP system. In order to assess the organizational transition, we evaluated the success of ERP systems implementation and measured the level of process-orientation before ERP systems implementation (phase 1), during ERP system Implementation (phase 2), and after utilization of ERP system (phase 3). Our goal was to understand

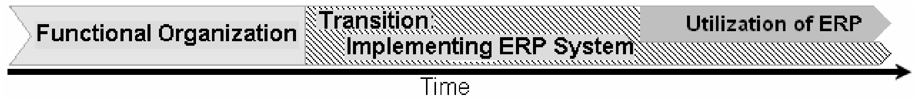


Fig. 2. Sequence of events in the transition from functional organizational structure to a process-oriented organizational structure

the relationship between the level of process-orientation in the three phases of the ERP system implementation and the success of ERP systems implementation.

The data was collected from 3000 Canadian and American large corporations through mail survey and online survey with an average total number of employees of 1,700, and average annual revenue of $4.8 billion. The questioner was sent through mailed and emails to the managers, CEOs and directors of the corporations. They were asked to participate in the survey if they had experience with implementing ERP systems in their organization. We received 275 responses.

This section consists of two parts. First we will discuss the measurement of process orientation in three phases, pre implementation of ERP system, during Implementation of ERP system, and post utilization of ERP system. In the following part the success of ERP system implementation and utilization is discussed.

4.1 Measurement of Process Orientation

We used our proposed system of process-orientation measurement. The respondents were asked to indicate the level of disagreement or agreement with about each of the 19 measures of process orientation in three phases. The questionnaire was designed to provide a 5 point Likert scale. We then calculated the average score of process-orientation in each of the three phases. We call this ERP Process-Orientation Score (EPOS). The findings indicate that the EPOS (level of process-orientation) increased from 2.53 in phase1, to 3.1 in phase2, and 3.4 in phase3.

For the purposes of data analysis we consider organizations with EPOS of less than 3 (median) not to be process-oriented and organizations with EPOS of equal or more than 3 to be process-oriented. Based on the findings from EPOS, organizations generally were not to be process-oriented in phase one. In phase two organizations in the sample became processes-oriented, and in phase three by utilization of ERP system, the level of process-orientation in firms increased even further. Based on the EPOS of the organizations in three phase of transition, three significant types of firms can be identified: Group1 (G1), Group2 (G2), and Group3 (g3). G1 are organizations that had EPOS of above 3 in the phase1 –and therefore they were process-oriented in the phase1– and their level of process-orientation increased in phase2 and phase 3. G2 firms are organizations that were not process oriented in phase1, but their EPOS increased to above 3, and they became process oriented in the phase2. Their level of process-orientation for G2 firms increased in the phase3. G3 are organizations that their level of EPOS was less than 3 in phase1. However, their level of process-orientation increased in the phase2 but it was still below 3. G3 firms could only become process oriented in the phase3. There were other groups of organization which their EPOS changed differently, however they did not count for a significant number of firms in our sample.

4.2 Assessment of ERP System Implementation

We assessed the two dimensions of ERP system implementation, success of implementation, and success of utilization. Each construct was measured using a number of measures (see Table 2 and 3). For each measure the respondents were asked to indicate their level of disagreement or agreement on a 5 point Likert scale. Based on the responses, we calculated the ERP System Implementation Score (EIS) and ERP system Utilization Score (EUS) for the three groups of firms. The higher EIS or EUS indicates higher level of success in implementing or utilization of ERP system. Table 3 presents the EIS across the three groups of firms. The data shows that in average all of the firms were successful in implementing and utilizing ERP systems with EIS and EUS of above 3. Table 4 displays the score of utilization of implementation of ERP systems across the three categories of firms. G2 firms were found to be the most successful firms in successfully implementing and utilizing ERP systems.

Table 4. Score of success of Utilization and Implementation of ERP systems

	G1	G2	G3
EIS	2.9	3.7	3.2
EUS	3.5	3.7	3.4

For better understanding the relation between EIS, EUS and process-orientation in organizations, a correlation coefficient analysis is conducted for three categories of firms. Figure 3 displays the correlations among process-orientation, success of ERP system implementation, and success of utilization of ERP system.

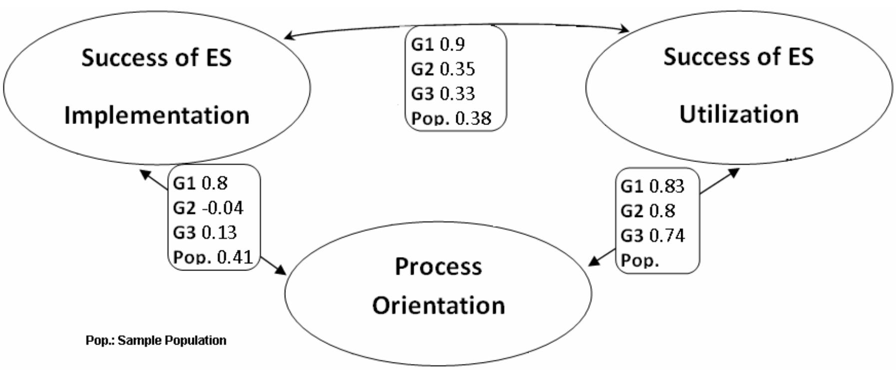


Fig. 3. Correlation among three constructs of ERP system implementation

5 Discussion of Findings

For understanding the transition process of each of the three groups the results of EIS, EUS as well as correlation analysis is explored. Our analysis of the EIS shows that G1 firms, in comparison to G2 and G3 firms were significantly less successful in "allocating sufficient time for ERP system implementation". "providing opportunity for continuous learning about ERP system", "paying enough attention to understanding and applying industry best practices for implementing ERP system", "putting in place systems to support, encourage and reward teamwork and team development", "paying enough attention to understanding and applying industry best practices for managing the deployment of ERP system" and "availability of shared information for decision making." For the above mentioned factors G1 firms received EIS of less than 3, while the G2 and G3 firms received higher EIS. G2 firms were found to be the most successful group of firms in successfully implementing ERP system. They were more successful than G1 and G3 firms on all of the success measures of ERP implementation. In comparison with G1 and G3 firms, G2 firms were significantly more successful in "allocating sufficient time for the ERP implementation", "utilizing ERP system in providing better access to its suppliers' and customers' information", "providing clear and well-defined roles for employees by management to eliminate resistance to change", and "the level of uncertainty among the employees about their involvement and role in the change process due to ES implementation". G3 Firms were found to be less successful than G2 firm in implementing ERP system. The G3 firms were significantly less successful in providing training for employees to understand and use end-to-end business processes, providing clear vision and well-defined role for employees to eliminate resistance to change, existence of ambivalence on how the organization manages IT investment, and lack of attention to people-based factors that would support the successful use of the deployed processes.

The analysis of the EUS of ERP systems also provides valuable information. G2 firms were found to be the most successful firms in using the ERP system to find "new sources of revenue", gain "clear financial benefits", achieve "cost saving", "access to more up-to-date and flexible infrastructure", and to help employees to have "more satisfying jobs". However G2 firms were found to be the least successful firms in delegating "autonomy in decision making" to the employees. G3 firms were found to be relatively the least successful firms in utilizing ERP systems. The main measures in which G3 firms score relatively less than other categories of firms were found to be, "flexibility and responsiveness in delivering products and services", "finding new sources of revenue", and "access of employees to higher quality data". G1 firms were less successful in utilizing ERP systems than G2 firms, however they were found to be marginally more successful in utilizing the ERP systems. The main factors contributing to the relative less success of G1 firms were found to be, "less success in providing flexible and up-to-date technical infrastructure", "little positive changes in jobs and individuals' roles", less success in gaining "financial benefits", and less success on achieving significant "cost saving" as a result of deployment of ERP system. The correlation coefficient analysis confirms the existence of positive relationship between the level of process orientation in organizations and success of utilization and success of implementation of ERP systems. However the correlation between the constructs of process-orientation and success of implementation of ERP system was not found to be at a significant level for G2 firms. Furthermore the correlation between the success of ERP system implementation and success of utilization of ERP system was found to be positive for all categories of firms.

6 Conclusion

During the ERP systems implementation organizations go through a transition in the way they perform their processes. The literature recommends that the organizations start the transition towards process-orientation and the process of ERP system implementation at the same time. However, most of the studies in this area use the case study approach, and as a result the findings can not be generalized. The current study using a large sample of North American firms provides an empirical base for this perspective. According to this perspective, organizations should start this transition with identifying organizational functions and current business processes, and subsequently design a refined business process based on the requirements of process-orientation structure. In the next step of the transition, the ERP system should be designed based on the refined business processes. During the implementation stage, there is a need for constant monitoring of the business process to identify the changes that can enhance the business. This is the same stage where paying attention to the factors contributing to the process-orientation of the organization, plays an important role in the success of implementing ERP systems.

The results of the study confirm that companies that organizations that become process oriented during ERP system implementation are not only more successful in implementing ERP system, but they will be more successful in utilizing and benefiting from the ERP system. We also explored the reasons as to why these firms are more successful in ERP system implementation and utilization. The findings indicate that G1 firms first redesign their processes based on their existing information platform and then implement the ERP system. In this situation, the G1 firms expect that ERP system adapts to the new processes institutionalized in the organization. These firms show less flexibility in changing processes in order to match with the ERP system compatibility requirements. On the other hand, G3 firms implement ERP system based on the current organizational processes, and after that, they start to change the processes within the organization to match with the requirements of the ERP system. At this time the newly implemented ERP system has relatively less capacity for modification, and as a result the processes are expected to adapt with the ERP system. Unlike the G1 and G3 firms, G2 firms have more opportunity to adapt their organizational processes with ERP system requirements, and also there is more opportunity for ERP implementers to modify the ERP system requirements with the processes in the organization.

References

1. Tchokogue, A., Bareil, C., Duguay, C.R.: Key Lessons from the Implementation of an ERP at Pratt and Whitney Canada. Int. J. Production Economics. 95, 151–163 (2005)
2. Balasubramanian, S., Gupta, M.: Structural Metrics for Goal Based Business Process Design and Evaluation. Business Process Management J. 11(6), 680–694 (2005)
3. Balzarova, M.A., Bamber, C.J., McCambridge, S., Sharp, J.M.: Key Success Factors in Implementation of Process-Based Management. Business Process Management J. 10(4), 387–399 (2004)
4. Biazzo, S.: Process Mapping Techniques and Organizational Analysis, lessons from Sociotechnical System Theory. Business Process Management Journal 8(1), 42–52 (2002)

5. Bingi, P., Sharma, M.K., Godla, J.K.: Critical Issues Affecting an ERP Implementation. Information Systems Management 16(3), 7–14 (1999)
6. Sheu., C., Chae., B., Yang, C.-L.C.-L.: National Differences and ERP Implementation. Issues and Challenges. Omega 32(5), 361–371 (2004)
7. Davenport, T.H.: Putting the Enterprise into the Enterprise System. Harvard Business Review, 121–31 (1998)
8. Ekanayaka, Y., Currie, W.L., Seltsikas, P.: Delivering Enterprise Resource Planning through Application Service Providers. J. Logistics and Information Management. 15(3), 192–203 (2002)
9. Fernandes, K.J., Raja, V., Autony, J.: Optimal Level of Goal Mapping in a Reengineering Environment. Business Process Management J. 7(1), 24–32 (2001)
10. Hammer, M.: Beyond Reengineering: How the Process-Centered Organization is Changing Our Lives. Harper Business, New York (1996)
11. Hammera, M., Champy, J.: Reengineering the Corporation: A Manifesto for Business Revolution. Harper Business Essentials, New York (2003)
12. Hendricks, K.B., Singhal, V.R., Stratman, J.K.: The Impact of Enterprise Systems on Corporate Performance: A Study of ERP, SCM, and CRM System Implementations. J. Operations Management. 25(1), 65–82 (2007)
13. Holland, C.P., Light, B.: A Critical Success Factors Model for ERP Implementation. In: IEEE Software., pp. 5–30 (1999)
14. Saraf, H.L., Xue, N.Q.H.Y.: Assimilation of Enterprise Systems: The Effect of institutional pressures and the mediating role of top management. MIS Quarterly 31(1), 59–87 (2007)
15. Llewellyn, N., Armistead, C.: Business Process Management: Exploring Social Capital within Processes. In: J. Service Industry Management., vol. 11(3), pp. 225–243 (2000)
16. Ljungberg, A.: Process Measurement. In: J. Physical Distribution and Logistics Management, vol. 32(4), pp. 254–287 (2002)
17. Lockamy, A., McCormack, K.: The Development of a Supply Chain Management Process Maturity Model Using the Concepts of Business Process Orientation. In: J. Supply Chain Management, vol. 9(4), pp. 272–278 (2004)
18. Love, P.E.D., Gunasekaran, A., Li, H.: Putting an Engine into Re-Engineering: Toward a Process-Oriented Organization. J. Operations & Production Management 18(9/10), 937–949 (1998)
19. McCormack, K.P., Johnson, W.C.: Business Process Orientation: Gaining the E-Business Competitive Advantage. St. Lucie Press (2002)
20. McQueen, R.: Work Culture and Process Improvement: Predictive Maintenance, Reno. In: NV: EPRIGEN Inc. (1999)
21. Mihailescu, D., Carlsson, S.A., Mihailescu, M.: Evaluating Enterprise Systems Implementation Methodologies in Action: Focusing Formalised and Situational Aspects. Electronic J. Info. Systems Evaluation 10(1), 83–90 (2007)
22. Fui-Hoon, N.F., Lee-Shang, L.J.: Critical Factors for Successful Implementation of Enterprise Systems. Business Process Management J. 7(3), 285–296 (2001)
23. Nah, F., Zuckweiler, K.M., Lau, J.L.: ERP Implementation: Chief Information Officers' Perceptions of Critical Success Factors. J. Human Computer Interaction. 16(1), 5–22 (2003)
24. Regev, G., Alexander, I.F., Wegmann, A.: Modeling the Regulative Role of Business Processes with Use and Misuse Cases. Business Process Management Journal 11(6), 695–708 (2005)
25. Sara, R., Aguilar-Saven,: Business Process Modeling: Review and Framework. International Journal of Production Economics 90, 129–149 (2004)

26. Shang, S., Seddon, P.B.: Managing Process Deficiencies with Enterprise Systems. Business Process Management J. 13(3), 405–416 (2007)
27. Vathanophas, V.: Business Process Approach towards an Inter-Organizational Enterprise System. Business Process Management J. 13(3), 433–450 (2007)
28. Vathanophas, V.: Business Process Approach towards an Inter-Organizational Enterprise System. Business Process Management J. 13(3), 433–450 (2007)
29. Zairi, M.: Measuring Performance for Business Results. Chapman & Hall, London (1994)

A Metric for ERP Complexity

Veena Bansal and Tripti Negi

Department of Industrial and Management Engineering
Indian Institute of Technology Kanpur, Kanpur 208016, India
{veena,negi}@iitk.ac.in
http://www.iitk.ac.in/ime/veena
http://www.iitk.ac.in/ime/PHD_IITK/negi.htm

Abstract. Enterprise Resource Planning (ERP) is a complex software. We use information flow metrics to quantify complexity of ERP software. The information flow metrics is defined in terms of flow of information within and among module/function/process. We compute complexity of sales and order process of a popular ERP package (SAP) using information flow metrics. We conclude that the complexity of the ERP package is due to the underlying business process. We then quantitatively show that configuration process of an ERP system is also complex. The complexity figures can be used to make the management commit appropriate resources to an ERP project. If an organization has used transaction processing systems, followed by management information systems and decision support systems in a networked environment, the organization may implement ERP using big bang approach. Otherwise, a phased approach is much safer and it will give organization time to mature.

Keywords: ERP, Complexity metrics, Configuration, Process complexity.

1 Introduction

An ERP system is an integrated enterprise computing system to automate the flow of material, information (data) and financial resources among all functions within an enterprise on a common database [11], [12]. An ERP system is a generic system that supposedly implements the best business practices. These best practices, the vendors claim, are the result of extensive research and reflect the practices followed by successful businesses across the globe. Many ERP solutions are available in the market such as SAP, Oracle, Baan etc. An ERP system makes an impact on technology, task, people structure, culture, strategy, size and environment of the organization. When an organization implements an ERP solution, data of all functional units gets integrated. The organization utilizes this data to achieve objectives such as reduce inventory, better production planning, better resource utilization etc. An ERP implementation takes anywhere from 1-3 years and costs 1-3% of the turnover of the organization. However, many of ERP implementation projects fail [4], [18], [21]. Almost 75% US companies experience some degree of failure in their ERP implementation [20]. An ERP system is a failure due to cost overrun, schedule overrun, system performance deficit

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 369–379, 2008.

or failure to achieve the expected benefits. Three main reasons for information technology (IT) project failures are poor planning or poor management (cited by 77%), change in business goals during the project (75%), and lack of business management support (73%)[20]. A project will suffer from poor planning, poor management or poor support only if the management does not understand criticality of their role. An ERP project is a business project but it is largely an IT project. An ERP package is a complex software and an additional reason for the failures [16], [21] of an ERP implementation is its complexity. It is hard to comprehend the complexity of an ERP implementation project, especially when ERP software is a semi-finished product that required only configuration and no programming. Configuration is a process whereby the individual components are assembled and adjusted to construct a working solution [3]. Components of an ERP software are relational database system that consists of thousands (30,000 in case of SAP) of tables (relations) and a large number of functions that operate on these tables. These components are assembled by selecting tables and their attributes; and setting up relationships among them. ERP system runs on top of an operating system that of course runs on top of computers that are connected by network. We will separate an ERP implementation into two sub processes: System Configuration: Select and configure appropriate hardware and software including database system, network system and operating system. Business Process Configuration (BPC): Configure ERP solution to capture business processes of the client organization.

System configuration is a technology process that needs to take database volume, number of concurrent users, availability (such as 24X7) requirements into consideration. In this article, we will not discuss system configuration further. Business process configuration captures the business processes. System configuration and business process configuration, both processes are fairly complex. Complexity is a well defined concept for software development and maintenance projects. In literature more than two hundred software complexity metrics have been proposed. Objective of these metrics is to assess complexity of an existing software product or the one that is being developed. Complexity is an important parameter in the models that estimate development or maintenance effort for a software project. Complexity is measured in terms of alternative paths loops, number of unique operators, operands, interactions of various modules etc. that exist in software. Interested reader may refer to Zuse's survey paper [22] on complexity metrics. In last 5 years, research community has shown some interest in benchmarking configuration complexity [2].The configurations actions may be captured in terms of canonical actions where each action is characterized by the number, type and source of its configuration variables. A recent article [6] discusses complexity of a business process using software complexity measures. The idea is that a business process can be seen as a software program where a set of lines of code or a function is replaced by a business process/activity. In order to compute complexity of a business process, it is modeled using one of the existing modeling techniques. Every modeling technique has a focus and one should use a modeling technique that has a matching focus with the objective of modeling. For example, Integrated Definition (IDEF0) and flow charts lay stress on activities and their relationships. These techniques give very little importance to the information flow [10]. There are methodologies such as data flow diagrams (DFD) that capture business processes through activities where data, information or any other intangible product or service and material flow from one activity to the other. It is observed that 70% or more of product flow within an organization (manufacturing or service) is information. This result confirms the claim by Toffler [19] that we are living in an information society and that organizations have become information organizations [8]. For a business process, data is important and a business process should be analyzed in terms of data and information that it modifies. Data is an important constituent of business processes and one needs to look at how data is defined, created, modified and used in an organization. In fact, enterprise resource planning software model an organization in terms of data and their relationships; and a set of functions that operate on the data. However, during the analysis phase before ERP implementation begins, an organization is modeled in terms of processes/functions using event process chain (EPC) or a similar modeling technique. Software complexity metrics have also been used to compute complexity of manufacturing control architectures [15]. A recent article [3] models a system that requires configuration as a set of containers and the configuration process as a chain of activities. For instance, application software that runs in Java runtime environment (JRE) is contained JRE; JRE runs on top of operating system that runs on top of the hardware. The application software in this case is a four level hierarchy. The focus of the paper is estimation of the configuration process complexity in terms of control flow and data flow (constructs from the software product complexity) complexity and an additional construct: memory complexity that captures the number of parameters that will have to be remembered by the individual who is configuring the system. It turns out that configuration process for a simple application has very high complexity. One can use complexity measure of the configuration process to tell the management and all other stakeholders the complexity of the seemingly simple configuration process. It turns out that the complexity of ERP software is because of the underlying data, their interaction and process complexity. We then use existing results from the literature to show that configuration of a complex system like ERP is complex. We have concentrated our effort on SAP as it is the largest ERP solution provider. SAP itself is a big package that consists of twelve modules and another four supporting modules that capture business processes of an organization. There are about 30,000 tables that store the business and transaction data. There are 3000 individual EPC (event process chain) models in SAP reference model. Since it is not feasible to investigate all these models, we decided to concentrate on sales and order process of SAP to illustrate our results. The complexity of a business process is related to the number of departments and activities involved. An activity becomes more complex as the flow of goods and information increases [10]. One can look at complexity as a characteristic of software interface that influences the resources another system will expand or commit while interacting with the software [7] or as the sum of structural and data complexity of the number of modules change [5]. In this paper we will look at complexity using the later view point. Rest of the paper is organized as follows. In the next section, we discuss our complexity metrics. In section 3, complexity of sales and order process of SAP is described and its complexity is computed. We conclude our paper in section 4.

2 Complexity Metrics

ERP configuration is a process whereby the individual components are assembled and adjusted to construct a working solution. Configuration requires no code to be written. The parameters that are set during the configuration process affect the data that will be available to each function/process, the data that can be passed between the functions/processes and the possible execution paths. We will model the process to be configured using EPC (event process chain) that captures control flow and data flow both. We use information flow metrics (influenced by the work of [9], [15] and [6]) and to compute complexity of the process that needs configuration. The information flow metrics is defined in terms of flow of information within and among modules (we use term module/function/process interchangeably). The flow of information can take place in three different ways as explained next. Local direct flow of information between two modules takes place when a module invokes another module and passes information to it or receives information from it (We do not discriminate between data and information). Local indirect flow takes place when invoked module passes the information obtained from invoking module to another module. A process can pass information to another process through global data structure that is termed as global flow of information. In an ERP system, data is stored in relational database system. The data thus is always global and a process passes information to another process through global flow. The information flow from and to a process is quantified in terms of fan-in and fan-out of the process. Fan-in is defined as the number of local direct, indirect and global flows into the process/module. The number of local and global flows that emanate from a module is called as fan-out of the module. Thus, inter-module complexity is defined as follows

$$Inter - module\ complexity = Fan - in + Fan - out \tag{1}$$

Intra module complexity is measured in terms of control flow complexity of the module. Control flow complexity is defined in terms of number of splits and joins. A decision point splits a path into alternative paths whereas alternative paths get joined at a joint into a single path. Process complexity is defined [15] as follows:

$$Process\ complexity{=}Intra{-}module\ complexity{*}(Inter{-}module\ complexity)^2 \tag{2}$$

A process may consist of many sub-processes. In order to compute process complexity, one can compute complexity of each of its sub-process and sum them up. We will now look at sales and order process of SAP and compute its information flow complexity.

3 Sales and Order Process and Its Complexity

We will use event process chain (EPC) to model sales and order process. In an EPC, events are passive elements. They describe under what circumstances a function works. Examples of events are *stock not available*, *quality test failed* etc. In the EPC graph, an event is represented as a hexagon. Functions are active elements in EPC. They model the tasks or activities within the company. Examples of functions are *create sales order, check stock* etc. In an EPC graph, a function is represented as rounded rectangle. A function in EPC diagram may end with more than one mutually exclusive events. In an EPC, there are objects such as information, material, or resource objects that portray objects from the real world. A function may use these objects as input or may produce them as its output. Examples are *material data*, *purchase order* data etc. In the EPC graph, an object is represented as rectangle. A control flow connects events with functions creating chronological sequence and logical interdependencies between them. A control flow is represented as a dashed arrow. Information flow shows connection of a function with input and output objects. Input objects are read and the output objects are changed or created. Information flow is also represented as an arrow. We have shown a small example in 1. An event order received triggers the function create sales order that creates the event sales order created. This event triggers the function check stock that may result into one of the two mutually exclusive events: stock available or stock not available. A function that procures stock will be triggered or sales will be made depending on the event occurred. We have shown sales and order process using EPC in figure 2. We have not shown events when only one event follows a function. This process creates a sales order for a particular customer after verifying his/her credit limit. Then the availability of the items is checked. If the items are available, the material is packed and dispatched updating the inventory. Customer invoice is created and the financial records are updated. If ordered items are not available, they have to be procured from the vendor by first creating a purchase requisition and then converting it to a purchase order. The purchase order will be given to a selected vendor who will supply goods. The received goods/items go through a quality check and finally goods receipt is created. Inventory is also updated. When invoice is received from the vendor, payment is made and necessary postings to the financial accounts are also made. The pending sales orders can now be processed. The functions used by sales and order process belong to materials management module (MM), sales and distribution module (SD) and financial (FI) module of SAP. These modules are virtual and used for explaining the functionality of SAP. Each function in our EPC accesses global data and communicates with other functions through global data. We are now ready to compute complexity of this process.

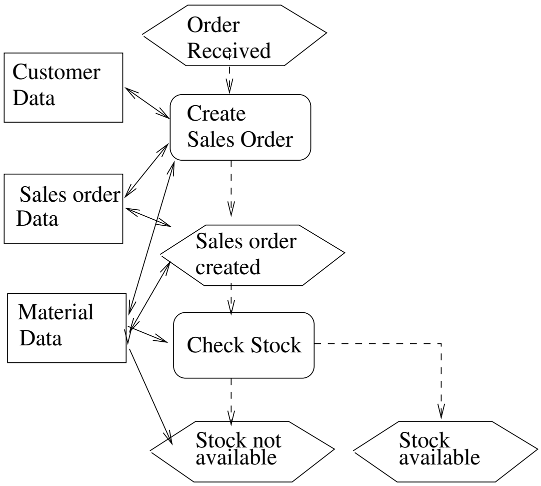


Fig. 1. An example of Event Process Chain; an event is shown as a hexagon, a function as a rounded rectangle, a control flow as a dashed arrow and an information flow as a solid arrow

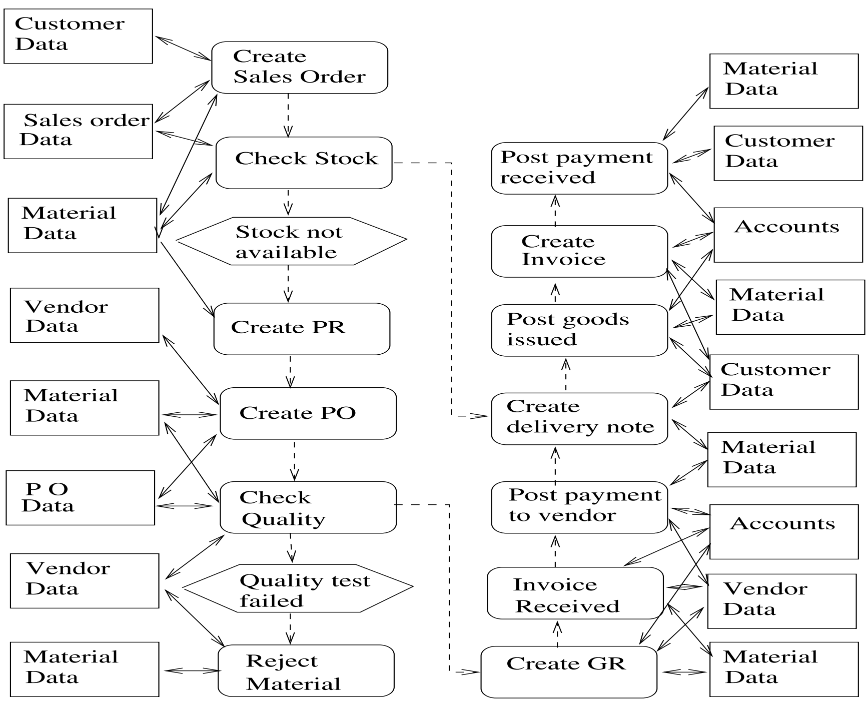


Fig. 2. Event Process Chain for Sales and Order Process of SAP; an event is shown as a hexagon, a function as a rounded rectangle, an object as a rectangle, a control flow as a dashed arrow and an information flow as a solid arrow

3.1 Inter-module Complexity Metrics for Sales and Order Process

Table 1 shows fan-in and fan-out for sales and distribution process depicted in figure 2. There are 15 functions in sales and order process. Each function reads and updates some global data as shown in 2. Customer data, sales order data, material data, purchase order data, vendor data and accounts data constitute global data for this process. Just to clarify further, create sales order function reads customer data, sales order data and material data in order to create a sales order. This function will update customer data, sales order data and material data while creating a sales order. These changes in the global data will be seen by subsequent functions. The fan-in and fan-out both are three for this function. We have computed fan-in and fan-out for each function and are shown in table1.

$$Inter-process\ Complexity = Fan-In\ +\ Fan-Out\ =\ 38+38=76 \quad (3)$$

Table 1. Inter-process Complexity of sales and order process shown in figure 2 in terms of fan-in and fan-out

Function	Fan-In	Fan-out
Create Sales Order	3	3
Check Stock	2	2
Check Credit (not shown in figure 1)	2	2
Tax Determination (not shown in figure 1)	2	2
Create PR	1	1
Create PO	3	3
Check Quality	3	3
Reject Material	2	2
Create GR	3	3
Invoice Received	3	3
Post Payment	3	3
Create DR	2	2
Post Goods Issued	3	3
Create Invoice	3	3
Post Payment Received	3	3
Total	38	38

3.2 Intra-module Complexity

The control flow complexity (CFC) of the sales and order process is computed in table 2. For a join, CFC count is one. That is, if two paths join at a point, the control flow complexity is 1. At a split, if a process is allowed to follow only one path from n alternatives, CFC is 2n-1. If the process is allowed to follow one or more paths from n alternative paths, CFC is n [6]. We have used all splits and joints present in sales and order process shown in 2 to compute its CFC. At a split, this process is allowed to take only one path. When there are two alternative paths, CFC becomes 3 using the formula $2n-1$ where n is 2. We can now compute the overall complexity which is

$$Intra{-}process\ complexity{*}(Inter-process\ Complexity)^2 = 21{*}(76)^2 = 121296$$

Our next job is to figure out the meaning and implication of this complexity figure: 121296. We looked at published literature and found that the high speed material transfer system developed at the university of Wisconsin Madison for real time delivery of tools, parts and raw materials in manufacturing systems has four modules and each modules has complexity in the range 10,000 to 22000 [15]. This system is considered complex by authors. We can safely conclude that *sales and order process* is a very complex process. Fan-in and Fan-out should not be more than 5 and number of branches in a process should not be more than 10.

Table 2. CFC Metrics for Sales and Order process shown in figure 2

Function	Split / Join	CFC
Create Sales Order		1
Check Stock	Split	3
Check Credit	Split	3
(not shown in figure 1)		
Tax Determination		1
(not shown in figure 1)		
Create PR		1
Create PO		1
Check Quality Split		3
Reject Material		1
Create GR		1
Invoice Received		1
Post Payment		1
Create DR	Join	1
Post Goods Issued		1
Create Invoice		1
Post Payment Received		1
Total		21

In an organization there are many such processes and the processes interact with each other. SAP R/3 includes all business processes of an organization that are complex and have complex interaction with other processes and data. One use of this metrics is to show the management that ERP system is a complex system. It is at times very difficult to convince the management that ERP software is a complex system. I cannot resist giving anecdotal evidence. In year 2006, Oracle approached an academic institute (320 faculty members and 2000 students) to sell their academic solution. A very senior person from the management told the Oracle team that we just want these three (pointing to one of their slides) bullet items; we have everything else. I was present at the meeting. He could not understand the fact that there will be a common database at the back of all these processes/bullet items and you just cannot have your own system and Oracle's bullets working together. It was even harder for Oracle to convince the top management of the institute that configuration will take couple of months. ERP community knows that SAP R/3 or any other ERP package provides a generic solution for the specific industry sector that needs configuration according to the specific needs of the organization and business. The configuration also is a complex process. We will use Brown's method and compute configuration complexity for SAP. We will use business process integration (BPI) one training module of SAP as our reference. This training module sets up a retail industry that operates from a single site and retails pens. The configuration requires:

– 60 actions
– Each action requires setting up 15 parameters on the average

The configuration process is easy on memory as SAP provides excellent cross referencing facility. Without distinguishing between usage of parameters (read, write, create etc.) and ignoring the context switches, the configuration is as complex as SPECjAppServer [3]. Authors see a need for creating automation tools for configuration of SPECjAppServer. Using these figures and references, we can easily convince the management about the complexity of SAP configuration process. Configuration requires experts that have in depth domain and process knowledge [1]. The existing effort and cost estimation tools for a software development project use complexity as a parameter along with many (as many as 15) other parameters. No such software yet exists for estimating ERP configuration cost. An organization benchmarks itself against an organization that has already implemented similar ERP system. It should be possible to use the complexity figure to compute configuration cost and implementation time for an ERP package. If an organization is not mature in terms of using information technology, an ERP system and its configuration may turn out to be too complex. An organization adopts information technology in four stages [13], [14]. The stages are initiation, contagion, control and integration. In the initiation stage, organization automates its well understood processes which fall into general category of transaction processing systems (TPS). The objective is to reduce functional cost. The organization would need only some knowledge of IT to achieve the objective. Only a handful of people will be involved and most of the employees would be unaffected. The management would need no justification for the IT spending as the expenditure is not much and the benefits start materializing quickly. In the second stage (contagion), more standard processes at operational level are automated that affect employees in general. The management continues to pump in money hoping to see better results. The employees in this relaxed environment become innovative. The users of information systems and information technology are no more only the IT people. People from other functional areas also start using IT. This relaxed environment doesn't continue for ever. The management puts control mechanism in place to start the third stage: control stage. The objective of the management is to achieve the better return on investments (ROI). By this time, however, the middle management realizes the potential of IT and starts using IT for decision making activities. Employees either assume or are assigned responsibility for IT applications. In the final stage, the organization starts integrating its IT applications that commences the integration stage. Users start taking active insert in IT. The management starts treating IT investments like other investments which go through cost-benefit analysis. ERP systems belong to the integration stage. If an organization has not gone through the first three stages, an ERP system will indeed turn out to be too complex. In fact, high percentage of ERP implementation failed in 80s and early 90s because many organization who opted for ERP implementation were not ready for ERP. We have shown quantitatively that even though ERP system requires no code to be written, its configuration is complex. An ERP implementation has better chances of success if role and function of people in steering committee, program managers, programmers, system managers and module implementation team from the client and consultancy company [17] are well understood and prioritized because ERP is a complex system.

4 Conclusion

It is hard for an ERP vendor to convince a potential client company to go for ERP system, specially when they have isolated IT system working. Once the client agrees to implement ERP system, It is even harder for it is even harder for the ERP solution provider to convince the top management that the configuration will take six months to one year when no programming/coding is required.

ERP system is complex due to the complexity of the underlying process whose configuration is equally complex. We have shown quantitatively that ERP is indeed complex software. We have used information flow metrics to quantify complexity of ERP software. The complexity figures can be used to make the management commit appropriate resources to an ERP project. No method or algorithm has been reported in literature to estimate implementation cost of an ERP system. We will see an algorithm similar to COCOMO to estimate effort for an ERP implementation in future as a result of further research. We should be able to use the complexity metric associated with an ERP implementation as one of the parameters to estimate cost and effort required in this algorithm.

References

1. Arinze, B., Anandarajan, M.: A Framework for Using OO Mapping Methods to Rapidly Configure ERP Systems. Communications of the ACM 46(2), 61–65 (2003)
2. Brown, A.B., Hellerstein, J.L.: An approach to benchmarking configuration complexity,Technical Report: RC23146(W0403-071) (March 10, 2004)
3. Brown, A.B., Keller, A., Hellerstein, J.L.: A model of configuration complexity and its application to a change management system. In: Ninth IFIP/IEEE International Symposium on Integrated Network Management, pp. 631–644 (2005)
4. Buckhout, S., Frey, E., Nemec, J.: Making ERP Succeed: Turning Fear into Promise. Strategy and Business, 2nd Quarter (1999) http://www.strategy-business.com/technology/99208/

5. Card, D., Agresti, W.: Measuring Software Design Complexity. Journal of Systems and Software 8(3), 185–197 (1988)
6. Cardoso, J.: How to Measure the Control-flow complexity of Web Processes and Workflows. In: IEEE International conference on Web Services (ICWS05), Orlando (2005)
7. Curtis, B.: Measurement and Experimentation in Software Engineering. In: Curtis, B. (ed.) Proceedings of the IEEE, 68th edn., pp. 1144–1157 (1980)
8. Drucker, P.F.: The new Realities. Harper & Row, New York (1989)
9. Henry, S., Kafura, D.: Software structure metrics based on information flow. IEEE Transactions on Software Engineering 7(5), 510–518 (1981)
10. Kock, N.F., McQueen, R.J.: Product Flow, Breadth and Complexity of Business Processes: An Empirical Study of Fifteen Business Processes in Three Organizations. Business Process Re-engineering and Management Journal 2(2), 8–22 (1996)
11. Kumar, V., Maheshwari, B., Kumar, U.: Enterprise resource planning systems adoption process: A survey of Canadian organizations. International Journal of Production Research 40, 509–523 (2002)
12. Motwani, J., Subramanian., R., Gopalakrishna., P.: Critical factors for successful ERP implementation: Exploratory findings from four case studies. Computers in Industry 56(6), 529–544 (2005)
13. Nolan, R.: Managing the computer resources: A stage hypothesis. Communications of the ACM 16(7) (1973)
14. Nolan, R., Croson, D.C., Seger, K.N.: The stages theory: A framework for IT adoption and organizational learning. Harvard Business School 9, 141–193 (1993)
15. Phukan, A., Kalava, M., Prabhu, V.: Complexity metrics for manufacturing control architectures based on software and information flow. Computers and Industrial Engineering 49, 1–20 (2005)
16. Rolland, C., Prakash, N.: Bridging the Gap between Organizational needs and ERP functionality. IEEE Requirements Engineering Journal 5(3), 1–25 (2000)
17. Sathish, S., Pan, S.-L., Raman, K.: Using Stakeholder Theory to Analyse Knowledge Sharing During Enterprise Systems. In: Implementations, 8th Pacific Asia Conference on Information Systems, Shanghai, China, pp. 336–349 (2004)
18. Stedman, C.: Failed ERP gamble haunts Hershey. Computer world 33(44), 1–2 (1999)
19. Toffler, A.: Powershift. Bantam Books, New York (1991)
20. Umble, E.J., et al.: Enterprise Resource Planning: Implementation and Critical Success Factors. European Journal of Operations Research 146, 241–257 (2003)
21. Vogt, C.: Intractable ERP A Comprehensive Analysis of Failed Enterprise-Resource-Planning Projects. ACM SIGSOFT Software Engineering Notes 27(2), 62–67 (2002)
22. Zuse, H.: Software Complexity: Measures and Methods. Walter de Gruyter, Berlin (1991)

Data and Information Quality Assessment in Information Manufacturing Systems

Mouzhi Ge and Markus Helfert

School of Computing, Dublin City University,
Dublin 9, Ireland
{Mouzhi.Ge,Markus.Helfert}@computing.dcu.ie

Abstract. Organizations are more and more concerned about the increasing data and information quality issues in their information (manufacturing) systems. These issues have caused various organizational problems such as losing customers, missing opportunities and making incorrect decisions. Recognizing these issues, one of the crucial aspects for organizations to sustain business growth and competitive advantage is to be able to assess data and information quality. However limited research has been done to investigate data and information quality assessment in information manufacturing systems. This paper proposes a model to assess the quality of two major information sources in information manufacturing systems: data stored in database and information products delivered to users. The proposed model is applied to an information manufacturing system and an example database. The research findings have shown that the poor quality of data found in example databases is correlated to the quality of information products perceived by users.

Keywords: Information quality, Information manufacturing system, Information quality assessment, Information quality dimension.

1 Introduction

In a broad spectrum of organizations, a number of business initiatives have been delayed or even cancelled, citing poor information quality (IQ) as the main reason. For instance, a major financial institution is embarrassed because of a wrong data entry of an execution order of $500 million (Wang et al. 2001). The explosion of space shuttle Challenger and the shooting down of an Iranian Airbus by the USS Vincennes are the results of IQ problems and IQ management errors (Fisher and Kingma 2001). From the cases above, we can observe that IQ problems are pervasive (Wang et al. 2001), costly (Eppler and Helfert, 2004) and even disastrous (Fisher and Kingma 2001). In order to prevent IQ problems, researchers have focused on heterogeneous aspects of IQ research such as IQ assessment and IQ management.

Most IQ research can be categorized into two research communities: database and management (Oliveira et al. 2005). The database community is typically technique oriented and usually uses the term "data quality", which is defined as the data that meet specifications or requirements (Kahn et al. 1998). On the other hand, the

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 380–389, 2008.

management community is business oriented and mostly refers to the term "information quality". Focusing on information consumers, information quality is defined as the information that is fitness for use by information consumers (Wang and Strong 1996). However, in a variety of studies (Wang and Strong 1996, Strong et al. 1997, Kahn et al. 2002), data quality and information quality have been used interchangeably. To include both views, in this paper, we use IQ to represent both quality of data stored in database and quality of information products delivered to users.

Recently, one novel perspective in IQ research is to observe information system as manufacturing system (Ballou et al. 1998). Three types of data are associated with this perspective: raw data, component data and information products (Batini and Scannapieco 2006, 7). Raw data are considered as raw materials for information manufacturing. They are typically well structured and stored in the database. Initiating information manufacturing, raw data are assembled into component data, which are then transmitted though different business manufacturing processes. Component data are used until information products are manufactured. Finally information products are delivered to users for indented use such as decision making and organizational management. Observing the above procedure, information manufacturing can be regarded as the process that transforms raw data to component data and finally to information products. While raw data and component data are usually stored in the database for manufacturing, information products are used for intended business operations. Therefore we can find two major sources of information: data, including raw data and component data, stored in database and information products delivered to users.

Based on the identification of information sources, the objective of this paper is to provide a comprehensive model for data and information quality assessment in information manufacturing systems.

This paper is structured as follows: section 2 provides a review on data and information quality assessment. Section 3 proposes a model which is used to assess data and information quality in information manufacturing systems. Based on the model, section 4 applies this model to an example dataset and its corresponding information products. Finally section 5 concludes this paper by providing research findings and future works.

2 Literature Review

A number of studies have confirmed that IQ is multi-dimensional concept (Ballou and Pazer 1985, Wand and Wang 1996, Wang and Strong 1996). Through the last two decades, different sets of IQ dimensions have been identified from both the database and management perspectives. Wang and Strong (1996) propose three approaches to study IQ, which are intuitive, theoretical and empirical approach. We adopt the three approaches to analyze IQ dimensions. Intuitive approach derives IQ dimensions from the researchers' experience and demands of particular cases. In this approach, the dimensions are identified according to the specific application contexts. For example, O'Reilly (1982) uses accessibility, accuracy, specificity, timeliness, relevance, and the amount of data to assess IQ in the context of decision making. Ballou and Pazer (1985) employ accuracy, timeliness, completeness and consistency to model IQ deficiencies in multi-input, multi-output information systems. The theoretical approach generates IQ dimensions on the basis of data deficiencies in information manufacturing process. For instance, Wand and Wang (1996) use ontology to derive IQ dimensions. They observe inconsistencies between real-world system and information system. The empirical approach captures IQ dimensions by focusing on whether the data are fitting for use to data consumers. For example, Wang and Strong (1996) provide 15 IQ dimensions that are important to data consumers. Kahn et al. (2002) use 16 IQ dimensions to deliver high quality information to information consumers. From the above analysis, we can observe that different IQ dimensions are identified using different approaches.

Adapting the definition of assessment (Gertz et al. 2004), IQ assessment can be defined as the process of assigning numerical or categorical values to IQ dimensions in a given setting. Based the identification of IQ dimensions, a variety of methodologies are proposed to assess IQ. These methodologies can be categorized into objective and subjective assessment (Pipino et al. 2002). Objective assessment methodologies use software to automatically measure IQ by a set of quality rules. They define the dimensions from database perspective and focus on the algorithms of IQ dimensions. For example, Cappiello et al. (2004) propose the algorithms of currency, accuracy and completeness in the context of multi-channel information systems. Subjective assessment methodologies use surveys or interviews to measure IQ by the information consumers. They define the dimensions from a management perspective and concentrate on whether the information is fitness for use. For instance, Lee et al. (2002) develop the AIMQ methodology to assess IQ by information consumers. Using IQ assessment, IQ can be determined in different organizational contexts such as accounting, decision making and data warehouse.

3 Research Model

Based on the literature review, we develop an IQ assessment model which consists of three major components: information sources, assessment methodologies and IQ dimensions. In the information manufacturing system, quality of raw data and component data are assessed by an objective IQ assessment method. Four dimensions are found to be widely used in objective IQ assessment. They are accuracy, completeness, consistency and timeliness. Quality of information products are assessed by a subjective IQ assessment method. Subjective IQ assessment uses a comprehensive set of IQ dimensions, which are defined from the perspective of data consumers. In our model, IQ dimensions are adopted from Wang and Strong (1996) that is one of the most cited and influential studies in IQ research. They identify 118 IQ items and use exploratory factor analysis to derive 15 IQ dimensions. To organize theses IQ dimensions for assessment methods, accuracy, completeness, consistency and timeliness are grouped into one category. Other IQ dimensions are grouped into another category. Based on the above discussion, we can develop the following research model described in figure 1.

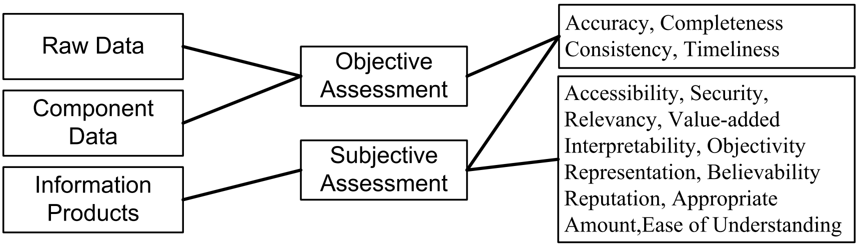


Fig. 1. Model for IQ assessment

To further clarify differences between objective and subjective assessment, we compare them from five aspects: tool, measuring object, criteria, process, assessing result and data storage. While objective IQ assessment uses data rules or patterns in the software to automatically assess the quality of data in database, subjective IQ assessment asks user to evaluate whether the information products are fitness for use. Table 1 summarizes the results of the evaluation.

Table 1. Comparison between objective and subjective IQ assessment

Method / Feature	Objective Assessment	Subjective Assessment
Tool	Software	Survey
Measuring Object	Data	Information Products
Criteria	Rules, Patterns	Fitness for use
Process	Automated	User Involved
Assessing Result	Single	Multiple
Data storage	Databases	Business Contexts

The advantage of using an objective assessment methodology is that large datasets can be automatically assessed. However it fails to capture the expectations of data consumers. That means high objective IQ may be not fitness for intended uses. For example, highly accurate and complete information may be very hard to understand. Using a subjective assessment methodology, IQ can be assessed as whether the information is fitness for data consumers' tasks. However it is difficult to assess large amount of information. Furthermore, since multiple assessment results are obtained from different information consumers, these assessment results need to be further coordinated and uniformed.

For objective assessment, we develop a strategy to assess the quality of data, which consists of 5 steps: (a) Making specifications. We specify objective values and place them into an individualized field in the database. (b) Identifying IQ problems. Based on the specifications, we recognize the problems which violated the specifications. (c) Linking IQ problems to IQ dimensions. If one IQ problem is connected to multiple IQ dimensions, it may cause the dependencies among these IQ dimensions. In order to clarify the distinct function of each IQ dimension, we link one problem to only one IQ dimension. However, one IQ dimension can be connected to different IQ problems. The dimension that is not linked to any IQ problems is dropped in the assessment. (d) Assessing the quality of data. This is an automatic procedure performed by the software. (e) Generating the IQ report. This report provides the basis for the IQ analysis and improvement. Figure 2 summarizes the above strategy to assess the quality of data.

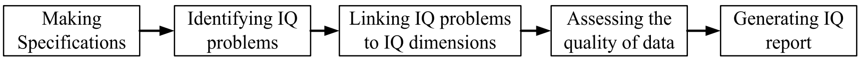


Fig. 2. Strategy for objective assessment

In the fourth step, the software performs the following assessing procedure: the software reads one data value and compares it to the specification. If the value conforms to the specification, the software will compare the next data value and so on. Otherwise the software classifies the inconformity into IQ problems and then records the corresponding IQ dimensions. After completing the assessment for the entire dataset, the software generates an IQ report. The overall procedure is illustrated in figure 3.

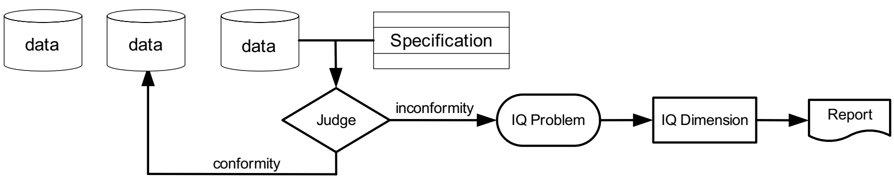


Fig. 3. Procedure of assessing quality of data

Wang (1998) proposes a total data quality management (TDQM) cycle to manage IQ. This cycle consists of four phases: define, measure, analyze and improve. Based on this cycle, the first three strategic steps are in the defining phase and the other two steps are in the measuring phase.

Following this approach, we develop a 5-step strategy for subjectively assessing the quality of information products: (a) Understanding intended use. As most IQ assessments are context related, users firstly have to understand the purpose of information utility. (b) Listing the information product inventory, according to the context, information products need to be identified. (c) Understanding IQ dimensions. Before evaluating IQ, users need to comprehend the definitions and subscales of each IQ dimension. (d) Assessing the quality of the information product. When completing the above three steps, users employ IQ dimensions to evaluate the listed information products. (e) Finally each user provides an IQ assessment report. Considering the TDQM cycle, the first three strategic steps are in the defining phase and the other two steps are in the measuring phase. Figure 4 provides an overview of the overall strategy.

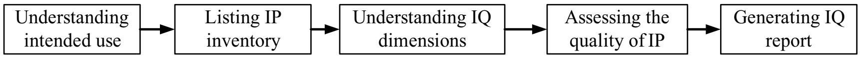


Fig. 4. Strategy for subjective assessment

In the fourth step, users evaluate the extent to which information products are fitness for intended use. The assessment procedure can be stated as users employ IQ dimensions to evaluate the quality of information products. This is shown in figure 5.

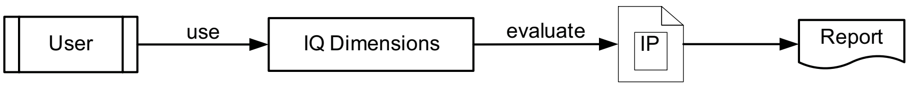


Fig. 5. Strategy for subjective assessment

Since subjective criteria and expectations vary from person to person, each user will generate an individual report. Different data consumers most likely generate different assessment results. For example, data custodians and managers draw completely different assessment results because some information is complete to data custodians but incomplete to managers. Therefore in the IQ analysis stage, we need to coordinate the subjective assessment results from several users or data consumers.

4 Application

The proposed research model was applied to an example dataset. The example dataset contained two objects that needed to be measured: data stored in the database and information products manufactured by the information system. In the example dataset we assumed that both data and information products were authorized to be accessed. This was necessary as information that is inaccessible, is meaningless to evaluate IQ. Without restricting our research, therefore we eliminated the dimension of accessibility and security in this applicatione

Using the strategy of objective assessment, we assessed IQ of ua_samsclub database in the Walton College Teradata system. Ua_samsclub database contained retail sales information gathered from sales at Sam's Club stores, which was a division of Wal-Mart Stores Inc. The database consists of 6 tables and 57 fields. Most fields are without specifications. The fields that provided specifications are centralized on the table member_index and store_visits.

For the purpose of our demonstration, we selected the table member_index as measurement object. In table member_index, 6 fields possessed specifications. Based on the given specifications, 4 IQ problems were identified and linked to different IQ dimensions. Timeliness was dropped because no IQ problem was connected to this dimension. We outlined each strategic step and action in table 2.

After the automated assessing procedure, a simplified IQ report was generated in table 3.

The report stated that 5668375 records were assessed and 119765 records contained IQ problems. As Sam's Club store is a membership-based store, information about its members is obviously crucial for their business. The results have shown that

Table 2. Objective IQ assessment

<table>
<tr><td rowspan="7">Make Specifications</td><td colspan="3">Field</td><td colspan="2">Specification</td></tr>
<tr><td colspan="3">BUS_CR_TYP_STAT_CD</td><td colspan="2">1-5,7,9</td></tr>
<tr><td colspan="3">CMPLMNTRY_CARD_CNT</td><td colspan="2">0,1,2</td></tr>
<tr><td colspan="3">ELITE_STAT_CODE</td><td colspan="2">0,2,3,4</td></tr>
<tr><td colspan="3">MEMBER_STATUS_CD</td><td colspan="2">A,D,E,T</td></tr>
<tr><td colspan="3">MEMBER_TYPE</td><td colspan="2">1,A,E,G,V,W,X</td></tr>
<tr><td colspan="3">QUALIFY_ORG_CODE</td><td colspan="2">null, 0015-3001</td></tr>
<tr><td>Identify IQ Problems</td><td colspan="5">P1. The data value is null except in the filed of QUALIFY_ORG_CODE.
P2. Figures are expressed by English such as describing 0 as zero.
P3. Spelling error such as case sensitivity.
P4. Except the above situations, data value does not conform to the specification.</td></tr>
<tr><td rowspan="5">Link IQ Problems to IQ Dimensions</td><td></td><td>Accuracy</td><td>Completeness</td><td>Consistency</td><td>Timeliness</td></tr>
<tr><td>P1</td><td></td><td>√</td><td></td><td></td></tr>
<tr><td>P2</td><td></td><td></td><td>√</td><td></td></tr>
<tr><td>P3</td><td>√</td><td></td><td></td><td></td></tr>
<tr><td>P4</td><td>√</td><td></td><td></td><td></td></tr>
<tr><td>Assess Quality of Data</td><td colspan="5">Automatic Procedure</td></tr>
</table>

Table 3. Objective IQ assessment results

Objective Assessment Report

Database: UA_SAMSCLUB, Table: MEMBER_INDEX, Records: **5668375**

	Accuracy	Completeness	Consistency
BUS_CR_TYP_STAT_CD	99.983% (933)	99.999% (17)	100%
CMPLMNTRY_CARD_CN_	99.999% (19)	100%	100%
ELITE_STAT_CODE	99.806% (10954)	100%	100%
MEMBER_STATUS_CD	100%	100%	100%
MEMBER_TYPE	99.833% (9418)	100%	100%
QUALIFY_ORG_CODE	98.263% (98424)	100%	100%

IQ deficiencies existed in the database. This may lead to many IQ related problems and possibly to high business losses.

To assess the quality of information products, we developed an electronic survey system to facilitate the evaluating procedure. The system consists of three major functions: The first function is to provide introductory information. The information includes a description of the software, procedure of identifying information products and definitions and subscales of IQ dimensions. The second function provides the evaluating environment.

Users evaluate the quality of information products by adjusting the slide bar of each IQ dimension. The slide bar is scaled from 0 % to 100 %. 0 % represented “not at all” and 100 % means “completely”. If the IQ dimension is not applicable for the current evaluation, users can label the dimension as “N/A”. The third function is to collect contextual information (e.g. intended use), demographic information and evaluation results. Based on the collected information, the software generates an assessment report for each user.

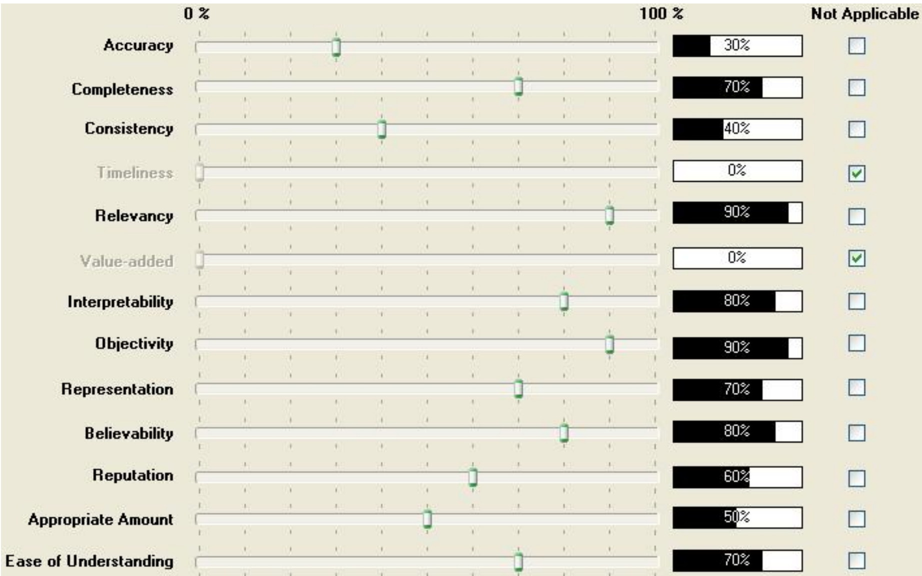


Fig. 6. Electronic survey system for subjective assessment

As a pilot study, the introductory information in the software was distributed to 10 researchers who had never used the database system. 7 researchers responded and 3 major comments were observed and detailed. Firstly, most respondents pointed out that the database needed to be specified because the database system contained different databases. Secondly, one respondent stated that concrete types of information products should have been illustrated, such as the usage of information or teaching information. Thirdly, two respondents suggested that system quality and information quality should have been differentiated, since users may consider whether the system worked well instead of just considering the IQ.

Considering the comments, we improved the software and carried out the evaluation with 20 system users (postgraduate students). These users employed the database system for different purposes at different levels. The evaluating results have shown that 75% of users assigned the accuracy and consistency as the lowest value. Thus we could observe that users were generally not satisfied with the accuracy and consistency of information products. 80% of users assigned relevancy as the highest value. A possible reason why relevance was rated high rated could be that users had filtered the irrelevant information products when listing the information product inventory. 25% of users assigned timeliness, value-added or ease of manipulation as “not applicable”. This indicated that the selection of the IQ dimensions should be according to the specific contexts. No user assigned 100% to all IQ dimensions. From the above results and analysis, we observed that in different contexts, IQ deficiencies existed in the example dataset. The subjective assessment of data quality conformed to the result of the objective assessment.

5 Conclusion

In this paper, we propose a model for IQ assessment in information manufacturing systems. The model consists of three major components: information sources, assessment methodologies and IQ dimensions. Information sources are raw data and component in the database and information products delivered to users. Assessment methodologies are divided into objective and subjective IQ assessment. Accordingly, two strategies are respectively developed for objective and subjective assessment. Based on the proposed model, we apply the strategies to an information manufacturing system. The research findings have shown that IQ deficiencies exist in this information manufacturing system. The identified IQ issues may result in significant losses in organizations, for example losing customers, missing opportunities and making incorrect decisions. Our research also indicates that the proposed assessment strategies provide a feasible prototype for practical IQ assessment. Furthermore, the model could be used for empirical deployment of IQ assessment in organizations.

In order to further investigate IQ assessments, three possible extensions of this paper are suggested. First, IQ needs to be assessed as a dynamic process or related business entities. In some information manufacturing systems, component data are operated and delivered between different processes, in which, the quality of component data can be amplified, diminished or kept unchanged. Thus determining the quality of component data is crucial for the quality control of information products. Second, after IQ assessment, IQ analysis is an important step for IQ improvement. IQ analysis is mainly to analyze assessment results and determine which data operations and organizational departments generate IQ deficiencies. Therefore IQ analysis provides the strategic foundations for IQ improvement. Third, estimations for costs of improving each dimension needs to be proposed. Considering the costs of poor IQ and the costs of IQ improvements, a model to determine the optimal total costs / benefits could be developed. This can assist organisations in improving their information quality.

References

1. Ballou, D.P., Pazer, H.L.: Modeling Data and Process Quality in Multi-input. Multi-output Information Systems. Management Science 31(2), 150–162 (1985)
2. Ballou, D.P., Wang, R.Y., Pazer, H.L., Tayi, G.K.: Modeling Information Manufacturing Systems to Determine Information Product Quality. Management Science 44(4), 462–484 (1998)

3. Batini, C., Scannapieco, M.: Data Quality, Concepts, Methodologies and Techniques. Springer, Heidelberg (2006)
4. Cappiello, C., Francalanci, C., Pernici, B.: Time-Related Factors of Data Quality in Multichannel Information Systems. Journal of Management Information Systems 20(3), 71–91 (2004)
5. Eppler, M., Helfert, M.: A Classification and Analysis of Data Quality Costs, Ninth International Conference on Information Quality, (November 5-7 MIT, 2004)
6. Fisher, C.W., Kingma, B.R.: Criticality of Data Quality as Exemplified in Two Disasters. Information & Management 39(2), 109–116 (2001)
7. Gertz, M., Ozsu, T., Saake, G., Sattler, K.: Report on Dagstuhl Seminar Data Quality on the Web. In: SIGMOD Report, vol. 33(1) (2004)
8. Kahn, B., Strong, D., Wang., R.Y.: Information Quality Benchmarks: Product and Service Performance. Communications of the ACM 45(4), 184–192 (2002)
9. Lee, Y., Strong, D., Kahn, B., Wang., R.Y.: A Methodology for Information Quality Assessment. Information & Management 40(2), 133–146 (2002)
10. Oliveira, P., Rodrigues, F., Henriques, P.: A Formal Definition of Data Quality Problems. In: Proceedings of the tenth International Conference on Information Quality, MIT (2005)
11. O'Reilly III, C.A.: Variations in Decision Makers: Use of Information Source: The Impact of Quality and Accessibility of Information. Academy of Management Journal 25(4), 756–771 (1982)
12. Pipino, L., Lee, Y.W., Wang, R.Y.: Data Quality Assessment, Communications of the ACM. Communications of the ACM. 45(4), 211–218 (2002)
13. Strong, D., Lee, Y., Wang, R.Y.: Data Quality in Context, Communications of the ACM. Communications of the ACM. 40(5), 103–110 (1997)
14. Stvilia, B., Gasser, L., Twidale, M.B., Smith, L.C.: A Framework for Information Quality Assessment. ournal of the American Society for Information Science and Technology 58(12), 1720–1733 (2006)
15. Wang, R.Y., Strong, D.M.: Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information System. 12(4), 5–34 (1996)
16. Wand, Y., Wang, R.Y.: Anchoring Data Quality Dimensions in Ontological Foundations. Communications of the ACM 39(11), 86–95 (1996)
17. Wang, R.Y.: A Product Perspective on Total Data Quality Management. Communications of the ACM 41(2), 58–65 (1998)
18. Wang, R.Y., Lee, Y.W., Ziad, M.: Data Quality. Springer, Heidelberg (2001)

A Comparative Study of Enterprise System Implementations in Large North American Corporations

Vinod Kumar, Bahar Movahedi, Uma Kumar, and Kayvan M. Lavassani

Sprott School Of Business, Carleton University,
1125 Colonel By Drive, Ottawa, Canada

Abstract. This study proposes measurement systems for assessing the success of implementation, challenges of implementation and success of utilization of enterprise systems. The proposed measurement systems are used empirically to assess the level of success and challenges of Canadian and American large corporations in the implementation of enterprise systems. Based on the findings of the study, a comparative analysis of the Canadian and American corporation is presented. The findings show that US firms are more successful in following their ES implementation master plan, in implementation of ES, and face fewer challenges. However, they are relatively less successful in utilizing the ES in comparison to Canadian firms.

Keywords: Enterprise System, ES Implementation, Challenges, Success, Utilization.

1 Introduction

The application of information technology tools in the integration of today's organizations is an inevitable fact. The Enterprise Systems (ES) that use multiple software and hardware modules are used to integrate processes and data in organizations. In this study ES is defined as an integrated, customized, and packaged modular-based system that includes compatible software and hardware and handles the majority of systems requirements in any or all of the functional areas of a firm. These areas include –but are not limited to– marketing, finance, human resources, and manufacturing. Based on this definition, almost every medium- and large-sized organization has at least several ES modules, such as a company-wide, accounting software package; a marketing software package; or a manufacturing software package. ES provides us with numerous promising functions – such as integration and automation of business processes, promoting common practices, sharing data across the organization, and providing real-time access to the information (Shari and Seddon, 2007; Fox, 2003; and Nah and Lau, 2001).

There is no doubt about the value that ES provides to any organization, however, the inability of some firms to successfully implement and utilize ES to increase

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 390–398, 2008.

organizational outcomes has been a source of concern for both practitioners and academia. The evidence of ES implementation failures go back to the late 1990s (Hayes, 2007; Hendricks, 2007; Davenport, 1998). In response to this, scholars in this field initiated a trend during the last one and a half decade to investigate the critical factors leading to successful ES implementation (Mihailescu et al., 2007; Huigang et al., 2007; Brown and Vessey, 1999; Holland and Light, 1999; Nah *et al.*, 2001; Nah, Zuckweiler, and Lau, 2003; Lee and Gosain, 2005; and Vathanophas, 2007). A vast number of studies done in regards to the success factors of ES are oriented towards case studies, and, as a result, they cannot be easily generalized; moreover the findings are usually limited to a specific area (Choi at al., 2007; Tchokogue *et al.*, 2005).

Therefore, in order to create a more global perspective, we designed a research study to explore different dimensions of ES implementation in North American companies. The present study examines large corporations in Canada and the US. It addresses the following subjects: the difference in the ES implementation, the level of success in implementing ES, the level and types of challenges that exist in implementing ES, and the degree of success in utilizing ES. For the purpose of this study we have developed four measures to help in assessing the implementation practice, the success of ES implementation, the challenges of ES implementation, and the success of ES utilization. The following section of this study presents the measures we developed that are based on our review of literature. The third section will discuss data collection and our analysis of the data gathered. The final section will present our discussion and analysis of the findings.

For better understanding the subject of this study, it is important to highlight the difference between the ES and the Enterprise Resource Planning (ERP) systems. The ERP system traditionally refers to a system that handles all activities of the firm from accounting and finance to managing human resources and customer relations (King and Burgess, 2006). While almost every company in developed countries uses one or several of these ERP modules, not many organizations have integrated all their operations under a global ERP system. Similarly, the evidence from the literature shows that, although many organizations are using some modules of an ERP system, they do not see themselves to be equipped with ERP (Keil and Tiwana, 2006; Rikhardsson, 2006; and Choi at al., 2007). Therefore, to circumvent this confusion we used the term ES implementation for the purpose of this study rather than the term ERP implementation, which implies to the implementation of all ERP modules.

2 Measurement of ES Implementation

Our first research goal in this study is to understand if the ES implementation is different in the US and Canada and, if so, what aspects are different. We measure three dimensions of ES implementation, namely, *success of ES implementation*, *challenges of ES implementation*, and finally *success of utilization of ES*. For assessing each construct we used a number of measures. In this section our proposed measurement system is described.

2.1 Measurement of Success of ES Implementation

Based on the review of the literature on the success factors of ES implementation, we developed a measurement system using 14 indicators. Each indicator measures one dimension of success in ES implementation. It is important to note that this measurement system does not deal with the effects of ES on the outcomes of utilization; rather it basically concerns the appropriate operation of the ES as it replaces the old system and promotes new processes. For each of the 14 factors the respondents were asked to indicate – based on their experience with ES implementation in their organization – the extent to which they agree or disagree with the presented success factor. We then developed the ES implementation success score (ESISS) with a numeric value between 1 and 5. ESISS is the average score of measures of ESISS. A high ESISS indicates a high level of success in implementing ES. Table 1 presents the list of factors that is used to measure the level of ES success in organizations.

Table 1. Measures of ESISS

1 Use of project management practices that effectively support the ES implementation	**9** Redesigning processes in conjunction with the ES implementation
2 Training most employees to understand and use end-to-end business processes using the ES	**10** Paying enough by the management attention to restructuring reward and incentive systems subsequent to ES implementation
3 Existence of systems to support, encourage, and reward teamwork and team development	**11**Enough attention to people-based factors that would support the successful use of the deployed processes
4 Existence of uncertainty among the employees about their involvement and role in the change process due to ES implementation	**12** Enough attention has been paid by the management to understand and apply industry best practices for managing the deployment of ES infrastructure
5 Readily shared information for decision making within the company	**13** Providing opportunities for continuous learning about ES functionality in the organization
6 Provision of clear vision and well-defined roles by management in order to eliminate resistance to change	**14** Allocation of sufficient time by the organization for the implementation of ES
7 Enhanced access to organization's, suppliers', and customers' information post deployment of ES	
8 Existence of ambivalence on how organization manages IT investments	

2.2 Measurement of Challenges of ES Implementation

For measuring the challenges of ES implementation, we developed an index named Challenges of ES Implementation Score (CESIS). This measure is the average score of on 13 measures, with each of them measuring a different dimension of the challenges of ES implementation. The list of measures developed, based on the literature review, is presented in Table 2. The respondents were asked to indicate their degree of agreement or disagreement with the existence of each challenge on a 5-point Likert scale. CESIS, which represents the average score of all challenges, is a number between 1 and 5. A higher CESIS indicates a higher level of challenges in ES implementation in the organization. The data gathered from this section strongly support the comprehensiveness of our proposed list of challenges, as the respondents hardly identified any other significant challenge than the ones used in the questionnaire.

Table 2. Measures of CESIS

1 The organization not making the right strategic choices needed to configure the systems and processes while implementing ES	**8** High turnover of key project persons
2 ES adoption decisions have not been viewed in terms of their strategic importance by top management	**9** Lack of a clear vision for the use of ES
3 In-house resource constraints	**10** Significant knowledge gap between implementers and users of ES
4 High costs of ES implementation	**11** Lack of sufficient training organized to provide employees with skills to use and maintain ES
5 Significant resistance from staff during ES implementation	**12** The chosen ES is inconsistence with organizational processes and does not complement organizational processes and policies
6 Existence of significant resistance from managers	**13** Difficulties in estimating ES project requirements
7 Lack of commitment from top leadership	

2.3 Measurement of Success of ES Utilization

To measure the success of utilization (outcomes gained) of ES, we also developed a list of factors based on our review of literature. For each item the respondents were asked to state their agreement or disagreement with the achievement of each success factor (outcome) resulting from ES implementation. The average score of these items is our Score of Success of Utilization of ES (SSUES). SSUES is the average score of measures of this construct, and therefore it is a number between 1 and 5. A higher SSUES shows a higher level of success in terms of outcomes of utilization of ES. The measures of SSUES are presented in Table 3.

Table 3. Measures of SSUES

1 Work has become easier	**9** Access to higher quality data and information about customers and suppliers
2 Employees have more collaboration with other employees in performing their tasks	**10** Access to a more up-to-date and flexible technical infrastructure post ES deployment
3 The jobs are more satisfying for the employees	**11** Major positive changes in the jobs and roles of individuals
4 Ability of the company to better pursue new business opportunities	**12** Employees have more autonomy in decision making that is directly related to their work
5 Clear financial benefits resulting from deployment of ES	**13** The implemented ES helps the company to reach and serve more customers than it previously did
6 Clear cost savings result from deployment of ES	
7 Creation of new sources of revenue resulting from deployment of ES	
8 More flexibility and responsiveness in delivering products and services across the organization	

3 Data Collection and Analysis

To collect data, 3,000 vice-presidents, directors, and managers of large US and Canadian corporations were contacted. The respondents were asked to respond to the survey if they have been involved in the implementation of ES in their organizations. The response rate was 9.1 percent (275 usable responses). The data was collected using 2,500 mailed surveys and 500 individualized emails. 48 percent of the respondents were Canadian firms and 52 percent were US firms. 3.5 percent of the firms were Canadian-American firms. On average, each company had 4.5 Modules of ES. The average for Canadian firms was 5.1 and for US firms it was 4.2. Finance and accounting modules, human resources modules, and supply chain management modules were among the most popular modules. By contrast, manufacturing, marketing, and project management modules were relatively less popular. Close to 30 percent of the firms indicated that they have some other ES modules that were specially designed for some specific functions. However, we could categorize each of those ES into one of our seven proposed ES modules.

For comparative analysis of ES implementation, challenges of ES implementation, and success of utilization of ES across Canadian and US firms, t-test is conducted for all measures of each construct (See Table 4). The last column in Table 4 shows the difference between the score of each measure for Canadian and US firms. Among the measures with the p-value of less than 0.05, we interpret the measures with larger difference t be more dissimilar across Canadian and US firms. In the following sections the comparative analysis of measures of each construct is described.

Table 4. T-test for Measures of Constructs of ES Implementation

Construct	Measures *	Mean of Measures		P-Value	Score of Canadian firms Minus the score of US firms
		Canada N=~120	USA N=~130		
Success of ES Implementation	1	4	3.9	0.245	0.12
	2	3	3.4	**0.000**	-0.41
	3	3.2	3.5	**0.000**	-0.34
	4	2.7	3.3	**0.000**	-0.60
	5	3.1	3.7	**0.000**	-0.59
	6	2.9	3.2	**0.013**	-0.35
	7	3.4	3.8	**0.000**	-0.40
	8	2.9	3.8	**0.000**	-0.91
	9	4.1	4.2	0.437	-0.07
	10	2.5	2.8	0.121	-0.21
	11	2.5	3.8	**0.000**	-1.23
	12	3.2	3.6	**0.000**	-0.44
	13	3.2	3.6	**0.002**	-0.35
	14	2.8	3.9	**0.000**	-1.06
Challenges of ES Implementation	**1**	2.5	2.5	**0.000**	0.69
	2	2.8	1.7	**0.000**	1.10
	3	4.1	4.1	**0.000**	0.98
	4	3.5	3.5	0.719	0.04
	5	3	2.5	**0.000**	0.46
	6	3	2.2	**0.000**	0.74
	7	2.6	1.8	**0.000**	0.85
	8	2.5	2.5	0.582	-0.07
	9	2.4	2	**0.000**	0.45
	10	3.3	3	**0.045**	0.29
	11	2.6	2.5	0.37	0.13
	12	2.2	1.9	**0.002**	0.33
	13	3.1	2.5	**0.000**	0.55
Success of ES Utilization	1	3.5	3.5	0.595	0.07
	2	3.2	2.8	**0.002**	0.38
	3	4.1	3.6	**0.000**	0.49
	4	3.4	3.8	**0.000**	-0.45
	5	3.7	3.7	0.283	-0.09
	6	3.8	3.8	0.787	-0.04
	7	3.2	3.1	0.515	0.11
	8	3.5	3.4	0.747	0.06
	9	3.9	4.2	**0.039**	-0.26
	10	3.6	4	**0.000**	-0.45
	11	3.7	3.7	0.705	0.03
	12	3	3.1	0.708	-0.04
	13	3.3	3.6	**0.009**	-0.31

* The numbers are associated with measures of each construct

3.1 Comparative Analysis of Success of ES Implementation

Higher ESISS among US firms indicates that Canadian firms are overall less successful in implementing ES. According to the analysis of measures, from the 14 measures of ES implementation, in 13 measures US firms received higher score, however only 11 measures were found to be significantly different across US and Canadian firms, at 90% confidence level.

The top five dimensions of ES implementation in which US firms in our sample were more successful are: attention to people-based factors that would support the successful use of the ES, allocation of sufficient time for ES implementation, ambivalent on how organizations manage IT investments, involvement and role of employees in the change process during ES implementation, and information sharing among different levels of organization for decision-making.

3.2 Comparative Analysis of Challenges of ES Implementation

Overall Canadian firms in our sample faced higher level of challenges with higher CESIS. For 9 measures of ES implementation Canadian firms faced significantly more challenges. The top five challenges in which Canadian firms were significantly scored higher than US firms are: ES adoption decisions not being viewed in terms of their strategic importance by top management, in-house resource constraint, lack of commitment from top leadership, resistance to change from managers, difficulties in estimating the ES project's requirements, and making the right strategic choices needed to configure the systems and processes.

3.3 Comparative Analysis of Success of Utilization ES

The overall ESISS score for North American firms in our sample was above 3 (3.4), which indicate overall success in implementing ES. However, while US firms achieved an overall ESISS of 3.6, the score for Canadian firms was around 3.1. US firms were found to be significantly more successful in 6 dimensions of ES implementation. In 4 dimensions of success of ES utilization, US firms were found to be more successful. The top two dimensions of success in which US firms were significantly more successful than Canadian firms are: pursuing new business opportunities, access to more up-to-date and flexible technical infrastructure. Moreover, for two dimensions of success of ES implementation, Canadian firms were found to be significantly more successful. These two factors are: to reach and serve more customers, and employees collaborate to perform their tasks.

3.4 Correlation Coefficient Analysis of Constructs of ES

The correlation coefficient among the three constructs of ES implementation is displayed in figure 1. Each correlation is calculated at three levels: among US firms, among Canadian firms, and among North American firms in our sample. However it is essential to note that for comparative analysis of correlations, more adequate statistical methods such as structural equation modeling is required. Using correlation coefficient we can merely conclude about the direction and strength of each correlation independently. The interpretation of correlation coefficients is done based on Cohen (1988) framework.

The correlation coefficient analysis indicates the existence of a positive correlation between success of ES implementation and success of utilization of ES for North American firms. Also for North American firms, higher level of challenges in implementation of ES was found to be correlated with lower level of success in implementation and utilization of ES. However, the relation between challenges of ES implementation and success of utilization of ES for US was found to be not significant. One reason for this low correlation coefficient can be the relative less challenges of US firms, couples with low superiority of US in successfully utilizing ES.

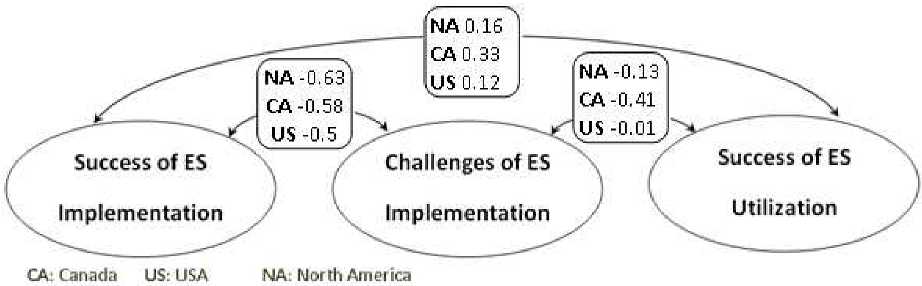


Fig. 1. Correlation Coefficient of Constructs of ES

4 Conclusions

This study explores the implementation of ES in large North American corporations. The relatively similar portion of Canadian and US firms in the sample data provides a good opportunity to perform comparative analysis of the data. The comparative analysis of the measures of the three main constructs of ES implementation, revealed significant differences between Canadian and US firms in various dimensions of ES implementation, challenges of ES implementation and in the utilization of ES. The findings indicate that US firms are relatively more successful in implementation of ES, face fewer challenges during ES implementation, while their success in utilizing ES has marginal difference with Canadian firms. Fur therefore the correlation coefficients analysis among the three major constructs of ES implementation is presented. However the comparative analysis of the correlation coefficients is limited to independent analysis of the correlations. In future studies the authors intend to use structural equation modeling to conduct advance comparative analysis on the relationship among the constructs of ES implementation.

References

1. Bingi, P., Sharma, M.K., Godla, J.K.: Critical Issues Affecting an ERP Implementation. J. Information Systems Management 16(3), 7–14 (1999)
2. Choi, J., Ashokkumar, S., Sircar, S.: An Approach to Estimating Work Effort for Enterprise Systems Software Projects. Enterprise Information Systems 1(1), 69–87 (2007)
3. Sheu, C., Chae, B., Chen-Lung Yang, C.-L.: National Differences and ERP Implementation. Issues and Challenges. Omega 32(5), 361–371 (2004)

4. Cohen, J.: Statistical Power Analysis for the Behavioral Sciences, 2nd edn. Erlbaum, Hillsdale, NJ (1988)
5. Davenport, T.H.: Putting the Enterprise into the Enterprise System. Harvard Business Review, 121–31 (1998)
6. Ekanayaka, Y., Currie, W.L., Seltsikas, P.: Delivering Enterprise Resource Planning through Application Service Providers. J. Logistics and Information Management. 15(3), 192–203 (2002)
7. Ekanayaka, Y., Currie, W.L., Seltsikas, P.: Logistics Information Management, vol. 15(3), pp. 192–203 (2002)
8. Fox, P.: The Art of ERP Done Right. Computerworld. 37(20), 3–22 (2003)
9. Greenbaum, T.L.: Moderating Focus Groups: A Practical Guide for Group Facilitation, Sage, London (2000)
10. Hayes, S.: Providing Enterprise Systems. Practical Accountant 40(2) (2007) SR11-SR11
11. Hendricks, K.B., Singhal, V.R., Stratman, J.K.: The Impact of Enterprise Systems on Corporate Performance: A Study of ERP, SCM, and CRM System Implementations. J. Operations Management 25(1), 65–82 (2007)
12. Holland, C.P., Light, B.: A Critical Success Factors Model for ERP Implementation. In: IEEE Software., pp. 5–30 (1999)
13. Huang, A., Yen, D.C., Chou, D.C., Xu, Y.: Corporate Applications Integration: Challenges, Opportunities, and Implementation Strategies. J. Business and Management. 9(2), 137–150 (2003)
14. Liang, H., Saraf, N., Hu, Q., Xue, Y.: Assimilation of Enterprise Systems: The Effect of Institutional Pressures and the Mediating Role of Top Management. MIS Quarterly 31(1), 59–87 (2007)
15. Keil, M., Tiwana, A.: Relative Importance of Evaluation Criteria for Enterprise Systems: A Conjoint Study. Information Systems Journal 16(3), 237–262 (2006)
16. Kim, Y., Lee, Z., Gosain, S.: Impediments to Successful ERP Implementation Process. Business Process Management J. 11(2), 158–170 (2005)
17. King, S.F., Burgess, T.F.: Beyond Critical Success Factors: A Dynamic Model of Enterprise System Innovation. In: J. Information Management., 26th edn., pp. 59–69 (2006)
18. Mihailescu, D., Carlsson, S.A., Mihailescu, M.: Evaluating Enterprise Systems Implementation Methodologies in Action: Focusing Formalised and Situational Aspects. Electronic J. Info. Systems Evaluation 10(1), 83–90 (2007)
19. Fui-Hoon, N.F., Lee-Shang., L.J.: Critical Factors for Successful Implementation of Enterprise Systems. Business Process Management J. 7(3), 285–296 (2001)
20. Nah, F., Zuckweiler, K.M., Lau, J.L.: ERP Implementation: Chief Information Officers' Perceptions of Critical Success Factors. In: J. Human Computer Interaction, vol. 16(1), pp. 5–22 (2003)
21. Rikhardsson, P., Kræmmergaard, P.: Identifying the Impacts of Enterprise System Implementation and Use: Examples from Denmark. In: J. Accounting Information Systems, vol. 7(1), pp. 36–49 (2006)
22. Shang, S., Seddon, P.B.: Managing Process Deficiencies with Enterprise Systems. Business Process Management J. 13(3), 405–416 (2007)
23. Sheu, C., Chae, B., Yang, C.L.: National Differences and ERP Implementation: Issues and challenges. Omega. 32(5), 361–371 (2004)
24. Tchokogue, A., Bareil, C., Duguay, C.R.: Key Lessons from the Implementation of an ERP at Pratt and Whitney Canada. Int. J. Production Economics 95, 151–163 (2005)
25. Vathanophas, V.: Business Process Approach towards an Inter-Organizational Enterprise System. Business Process Management J. 13(3), 433–450 (2007)

Enterprise Resource Planning (ERP) Systems Implementation Challenges: A Kenyan Case Study

Jim Odhiambo Otieno

School of Business and Computing Science, Middlsesex University
Burroughs Road, NW4 4BT, London, UK
J.Otieno@mdx.ac.uk

Abstract. Enterprise Resource Planning (ERP) systems have transformed the way organisations go about the process of providing information systems. They promise to provide an off-the-shelf solution to the information needs of organisations. Despite that promise, implementation projects are plagued with much publicized failures and abandoned projects. The study investigates the challenges faced by organisations implementing ERP systems in Kenya. The findings of this study should provide to management of firms implementing ERP systems a better understanding of the likely challenges they may face and put in place appropriate measure to help in mitigating the risk of implementation failures.

Keywords: ERP, Challenges, Critical Success Factors, Implementation, Kenya.

1 Introduction

An Enterprise Resource Planning (ERP) system is an integrated set of programs that provides support for core organisational activities such as manufacturing and logistics, finance and accounting, sales and marketing, and human resources. An ERP system helps the different parts of an organisation share data and knowledge, reduce costs, and improve management of business processes. In spite of their benefits, many ERP systems fail (Stratman and Roth, 1999). Implementing an ERP system is a major undertaking. About 90% of ERP implementations are late or over budget (Martin, 1998) and the success rate of ERP systems implementation is only about 33% [1,2].

Over the past few years, limited research has been conducted about ERP implementation issues: mainly case studies in individual organisations have been reported. A major problem with such ERP case studies is that very few implementation failures challenges resulting to these failures have been recorded in the literature, and thus the reasons why implementations fail are not known to practitioners and researchers. That is a motivation toward conducting empirical studies to explore challenges that affect ERP systems implementation. In

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 399–409, 2008.

the context of ERP project implementation, challenges represent major pitfalls which if not addressed then a project stands little chance of success. This study examines the challenges faced by user organisations implementing ERP systems in Kenya. Managers from five user organisations, who were identified as having a key role in ERP systems implementation, were interview in stage 1 of the research in order to empirically identify challenges faced during ERP implementation. A survey was then conducted to assess the criticality of the identified challenges which were perceived by respondents as posing high risk to successful ERP implementation.

This paper is organized in four sections. First ERP-related literature is reviewed. The next section introduces the research methodology, followed by the presentation of the results. The paper ends with the conclusions and implications for future research and practice.

2 Literature Review

Lured by guarantees of improved business productivity, streamlined business operations, and increased cost savings, organisations worldwide have launched initiatives to integrate ERP systems into their existing business environments. There has been a growing increase in using Enterprise Resource Planning (ERP) systems developed by, for example, SAP, Oracle, BaaN, PeopleSoft and JD Edwards as a business information system platform for large organisations and government bodies in developed countries such as USA, UK, Canada, and Australia [3]. [4] states that:

> ERP systems have now been adopted by the majority of the Fortune top 500 firms, and as the high end of the market becomes saturated, ERP systems are filtering down to medium-sized organisations, and to regions beyond those initially penetrated in Europe and North America.

While there is a wide adoption of ERP system in Europe and North America, developing countries lag far behind. However, due to economic growth, developing countries such as Kenya are now becoming major targets of ERP vendors [5,6,7,8,9]. In some developing countries, for example in Kenya, a number of large and mid-sized organisations has implemented ERP solutions and more are expected to follow suit. The majority of adopting organisations that have joined the 'ERP bandwagon' [5,10] have presumed that with relative ease they can benefit from the alleged 'best-of-suite solutions' that are embedded within the business processes of these generic packages.

However, the transfer of information systems like ERP - typically developed in developed countries - to developing countries is often marred by problems of mismatch with local, cultural, economic and regulatory requirements. For instance, [9, p. 219]

> Meanwhile, ERP is beginning to appear in many organisations of developing countries. Little research has been conducted to compare the implementation practices of ERP in developed vs developing countriesERP

technology faces additional challenges in developing countries related to economic, cultural, and basic infrastructure issues.

In spite all the benefits implementing ERP is a risky undertaking. The truth is that due to the behavioural and management related challenges in the implementation process many ERP projects have been terminated. The reasons being: end-user not being ready, resistance to change, lack of user education and training, high turnover of key personnel, lack of communication and support documentation, the layer of consultants in addition to pure technological problems such as software bugs and configuration difficulties [11]. In summary, several studies agree that the biggest obstacles are people, organisational issues and change management [2,6,12,13,11,14]. Moreover, people challenges are considered to be more difficult to manage than the technical problems [11,15,16,17].

Popular press and trade journals have documented both stellar successes and miserable failures [18], but with very little explanation on the underlying causes. Since the trade press is now a replete with articles on ERP failure [5,19,17] in Europe and North America where most these systems originate from, we agree with [9] argument that its implementation is likely to be more problematic in less developed given that "ERP technology faces additional challenges in developing countries related to economic, cultural, and basic infrastructure issues". This is exacerbated by the claim that ERP embodies established ways of doing business thereby requiring organisations adopting ERP system to change their business processes to conform to business practices inbuilt in ERP packages.

There is an urgent need for understanding ERP adoption and usage issues in less developed countries and in Kenya in particular because these systems are still in their early stages in these countries and face economic, cultural and infrastructure challenges. The adoption and use of ERP in Kenya raises a unique question for less developed countries within Sub-Saharan Africa that has not been addressed in previous studies, i.e. the challenges faced by user and vendor organisations during ERP implementation. Because of the complex and integrated nature of ERP, and the large investment involved, it is imperative for organisations to be aware of the challenges and the experiences of others, and to learn from their challenges and practices [6,2,20]. We suggest that identifying challenges relevant to local companies is one way to increase the chances of a successful local ERP implementation.

3 Research Methodology

The choice of an appropriate research methodology is critical in guiding researchers on how best to meet research objectives. In this study, the purpose was to discover the perceptions and experiences of companies using ERP systems in Kenya and to use that information as the basis of data collection.

The target of the study was the organisations that have implemented ERP systems in Kenya. The key informant method was used for collecting information in a social setting by surveying (or interviewing) a selected number of participants. Five user organisations participated in the study. We contacted the ERP project managers of each company in charge of ERP implementation. About one hundred questionnaires were sent to the ERP project managers of each firm, who forwarded the questionnaires to the project team members in charge of individual processes. A total of 69 questionnaires were returned, of which 51 were valid.

The questionnaire consisted of two main parts: the company background and statements which expressed the challenges identified in stage 1 one of the study. The first part was designed to determine characteristics such as size of the company, type of industry, location of company etc. The second part consisted of nineteen statements derived from the literature review, expressing challenges facing ERP systems implementation. The language used in the survey was both English and Kiswahili (national language in Kenya). Translation was rather easy because Kenyans use original English terms for many technical and management concepts and especially for information systems and computing concepts.

Participants were requested to agree or disagree on a five likert scale with the statements which expressed the challenges identified during the interview stage. This method was employed on the grounds that it avoids the problems of having to consider nineteen challenges simultaneously in order to rank them. The data collected was then analyzed by using SPSS. Based on the responses, descriptive statistics, factor analysis (FA) and reliability tests were carried out to identify the critical challenges faced by user and vendor organisations during ERP implementation in Kenya and data validity respectively.

4 Results

Rankings The importance rating of nineteen challenges is listed in Table 1

The individual mean value of the Likert rating scale is the popular usage indicator for measuring an item's importance, without regard to the other items: so the higher the value the more important the factor. Most items are rated above the 3.0 scale (mid-point). The three most critical challenges, in order of declining importance, are: lack of skills, insufficient training, and incompatibility with work practices, with a mean value ranging from 4.99 to 4.83. Conversely, System taking too long to operate, vendors' unreliability, and poor ERP quality are the three items lowest in the list, with a mean value ranging from 4.17 to 3.17.

4.1 Factor Analysis

In an attempt to reduce the number of item (challenges), and to understand their underlying structure, a factor analysis (FA) was performed. FA is a data reduction technique that uses correlations between data variables. The underlying assumption of FA is that a number of factors exist to explain the correlations or inter-relationships among observed variables [21].

For the present study, FA was performed on all nineteen variables using principal components extraction [21]. The goal of this method is to extract maximum variance from the data set within each factor. It is basically used to reduce a large number of variables down to a smaller number of components. The measure

Table 1. Rankings

Challenge	Mean
Lack of skills	4.99
Insufficient training	4.94
Incompatibility with work	4.83
Data conversion problems	4.81
High system cost	4.71
Long customization period	4.68
Integration problems	4.67
Benefits not recognised	4.61
High user resistance	4.56
Inadequate preparation by employees	4.50
High staff turnover	4.34
Organisational changes	4.33
Complexity of ERP	4.29
System insecure	4.23
Inadequate implementation time	4.21
Staff layoff	4.20
System taking too long to operate	4.17
Vendors unreliability	4.10
Poor ERP quality	3.17

of sampling adequacy for the twenty items was 0.87 indicating that the items were suitable for factoring (Kaiser, 1974).

A three-stage factor analysis was conducted with an orthogonal (varimax) rotation to obtain a stable factor structure, resulting in easily interpretable factors. Under this three-round factor analysis, items were omitted according to the following two criteria: (1) no loading greater than 0.45, or (2) loading greater than 0.45 on two or more factors [22].

Table 2 shows the results of this analysis. A first factor analysis was conducted and produced five factors. According to the two criteria, two items were dropped. A second factor analysis on the remaining 17 items resulted in six factors and the dropping of one item. Finally, a three-factor structure was derived which kept a total of 16 items after three iterations. The minimum eigenvalue from a varimax rotation for which a factor was to be retained was set at 1.0 in order to satisfy the minimum eigenvalue criterion [23].

Factor 1, named "Integration and incompatibility issues", comprises three items relating to integration related issues such as data and file conversion, and work practices incompatibility: incompatibility with work, data conversation problems, and integration problems.

Benefits of an ERP application are limited unless it is seamlessly integrated with other information systems. Organisations face many challenges relating to ERP integration: (1) the challenges of integrating various functional ERP modules, (2) the challenge of integration with other e-business software applications, (3) the challenge of integration with legacy systems. Organisations noted that legacy systems have accumulated vast amount of data vital to the survival and operations.

Integration of ERP systems with legacy systems is more complex than the integration of ERP modules and Integration of e-business applications. It routinely requires the installation of third-party interface software for communication between ERP software systems and legacy systems. Second generation ERP systems use relational database management system (RDBMS) to store enterprise data. Data conversion from legacy systems to RDBMS is often a time-consuming and tedious process. Integration of the business processes also faced additional challenges related to new rules built into ERP software being incompatible with the established ways of thinking and the norms of behaviour embedded in the existing work routines. This is consistent with the idea of 'best practice' being situated. Assistant Purchasing in one of the organisations while explaining that ERP could not accommodate their work practice said that:

> Given our unique requirement, Ebizframe could not meet our need. The system could not accept advance payments. It required us to raise DN [Delivery Note] first then raise sales invoice and they receive payment against the invoice. We don't work that way.

Factor 2, named "High cost further escalated by extensive customization", comprises two items relating to ERP system cost: High system cost and long customisation period. Cost is a major impediment to the implementation and use of ERPs. Both the respondents in organisations that were interviewed alluded to this fact. The cost of ERPs implementation averages between Kshs. 100 to 500 million (USD 1.4 - USD 7 million). However, the cost of ERP implementation varies from one ERP to another. Baan and SAP R/3, two of the ERP systems implemented by case organisations, charge software license fee charged based on: 1) the number users, 2) the type of users, 3) the number of master records in the database. Users are categorised based on what they can do or not do in the system.

Cost continues to be a concern for many organisations. In the case of SAP R/3, their costing model seemed not favour Africa. For example, the cost charged per day per consultant is USD 1,200 plus and an additional hotel and overnight stay allowance of USD 400 charged per night. This is three times above the amount charged in Asian continent like in India that is USD 600 per day. Implementation costs were found to be, on average 25% percent over budget. Organisations under-estimated support costs for the year following initial implementation by an average of 20%. IS maturity had a major influence on support cost for the year following initial implementation. Organisations with low IS maturity experienced an increase in support cost whereas IS mature organisations experienced a decrease in costs. This disparity can be traced, in part, to the widely varying maturity of pre-ERP environments. Lack of regional standardization and low budget for IT within Kenya and other African nations makes it difficult for ERP companies to find markets with enough potential to justify investing in costly customisations of the products.

Factor 3, named "Poor change management and failure to realize ERP benefits", comprises four items relating to organisational changes triggered by ERP, resulting user resistance, and lack of realisation of benefits once ERP is operational: inadequate preparation by employees, organisational changes, benefits not recognised, and high user resistance.

Implementing an ERP will bring in changes to the way people work within the organisation, processes will change and there may be job cuts and rationalization of responsibilities within departments. All this will definitely evoke resistance from the employees and this has to be managed effectively before, during and after the implementation of the ERP package. Respondents noted that they were inadequately prepared for the major undertaking of ERP implementation. Most respondents noted that they were neither aware of the reasons for the implementation nor the business benefits that can be expected by implementing a new ERP package. ERP projects are therefore perceived as cost cutting measure by organisations which lead to mass staff layoff. Respondents noted that there was no strong change management team in place to oversee to approve, implement and track the changes in the organisation, which includes the impact and detailed structure (i.e. documentation) associated with the life cycle of the ERP project.

Factor 4, named "Unreliability of vendors and poor quality of some ERP systems" comprises two items relating to poor services by ERP vendors and consultants, and some ERP systems being substandard: vendors are unreliable and poor ERP quality. Respondents noted that the quality of some ERP systems is not to standard hence not giving many users the desired results as they are inflexible and lack some functionality. This factor was more prevalent with organisations whose ERP systems had not matured or were relatively new such as ORION and EBIZFRAME. Organisations which were running SAP R/3, Baan and Oracle Financials disagree with the statement that expressed this factor.

The respondents also felt that the vendor representatives were unreliable and lacked knowledge of ERP. Most of the consultants were always communication with their counterparts in the vendor headquarters using telephone conference to try to implement the modifications and customisations. The communication cost is meant to be catered for by the implementing company and is part of contractual agreement. One obstacle was that the people who were available in Kenya did not understand the software. IT manager interviewed described the lack of experience on the part of a consultant in Kenya as follows:

> The consultant in Kenya had limited experience. He always contacted the headquarters in India and held conferences. After 3 years of hard work without any result, we got upset. The vendor company then sent a new consultant which meant changes in the Implementation team.

The lack of confidence in knowledge of the consultant respondents noted often leads to ERP failure. The use of foreign experts in implementing the ERP software as well as the reliance on their headquarters abroad often leads to lack of confidence in ERP consultants in Kenya. Therefore vendors suffer from a lack of local expertise in their software.

Factor 5, named "Lack of skills by both users and high staff turnover" comprises three items relating to lack of capacity to cope with ERP on the part of all organisational members at all levels due to the lack of training and high staff turnover: insufficient training, high staff turnover, and lack of skills by users.

An all common complaint was the frequency with which the case study organisations lose key personnel experienced with ERP or supporting technologies. Frequently reported problems were: (1) losing key IT specialists and user representatives working on the project while the project was going on, often despite handsome retention bonuses, (2) losing experienced people after the project was complete. Many IT specialists thrive on project work and view assignment as a 'competence centre' and springboard to lucrative opportunities.

One of the major challenges facing ERP systems implementation in Kenya is the non-existence of well-qualified employees in implementing organisations to manage the implementation process of the system. In one of the case organisations, the ERP project was supervised by the financial controller (the ERP Project Manager) and the heads of administrations (the key users). None of them had any knowledge about the computer or the ERP software. The organisations use India and South Africa as resource base for implementation. The MIS General Manager said:

> All of the company's leaders were not qualified to use the computerThey only trained on the beginnings of Windows and DOSTraining was internal in finance department for one week. One week was not enough. It was just background information.

Factor 6, named "Complexity of ERP system further compromising its security" comprises two items relating to complexity of ERP and security threat exposure in using ERP: complexity of ERP and system insecure. Complexity refers to the degree to which an ERP is perceived as being difficult to understand and use. Most respondents agreed that ERP system complexity discouraged its adoption and lead to greater difficulty in its implementation and further usage. The ERPs are too complex for most users and thus users need to be trained adequately on how to use them. The implementation of ERPs is quite an involving exercise that calls for co-operation between the vendors, the project teams and the management. One respondent commented that:

> ERP has been notorious for its complexity. For example, SAP R/3 has more than 3000 configuration tables. We spent more than a year on only going through these tables

Respondents noted the complexity of ERP systems also leads to security vulnerabilities. ERP systems must be able process a wide array of business transactions and implement a complex security mechanism that provides a granular-level access to users. For example, in SAP R/3, hundreds of authorization objects are used to allow access to various actions in the system. A small or medium sized organisation may have 100 transactions that are commonly used, and each transaction typically requires at least two authorisation objects. If the company has 200 end users who fill a total of 20 different roles and responsibilities, there are approximately 800,000 (100*2*20*200) ways to configure security in the ERP system- and this scenario excludes other complexity factors, such as multiple transactions sharing the same authorisation objects, an authorisation object having up to 10 fields that can be assigned various values, and the possibility of using position-based security.

Cronbach alpha coefficients were calculated to test the reliability of these challenges, as shown in the last row of Table 2. The reliability of coefficients obtained ranges from 0.58 (factor 6) to 0.76 (factor 1). Srinivasan (1985) proposed that a coefficient of 0.7 or higher is acceptable, while a coefficient of 0.5 or higher is considered sufficient when dealing with exploratory research combined with unvalidated data. Thus, the reliability coefficients in this study are deemed acceptable. The strength of factor analysis is that it provides a basis for data reduction. Therefore rather than looking at all nineteen items, just six factors can be examined. That simplifies the rankings and clarifies the most important items. Rather than focusing on individual items, practitioners and researchers can focus on the broad set of items represented by the essential factors.

Table 2. Results of factor analysis

Challenge	Mean	F1	F2	F3	F4	F5	F6
Lack of skills	4.99					.792	
Insufficient training	4.94					.682	
Incompatibility with work	4.83	.689					
Data conversion problems	4.81	.610					
High system cost	4.71		.500				
Long customization period	4.68		.733				
Integration problems	4.67	.651					
Benefits not recognised	4.61			.591			
High user resistance	4.56			.912			
Inadequate preparation by employees	4.50					.792	
High staff turnover	4.34					.792	
Organisational changes	4.33			.699			
Complexity of ERP	4.29						.815
System insecure	4.23						.678
Staff layoff	4.20					.792	
Vendors unreliability	4.10				.814		
Poor ERP quality	3.17				.790		
Eigenvalue		3.52	3.00	2.05	1.63	1.45	1.24
Percentage of variance		18.5	15.54	10.78	8.56	7.61	6.52
Cummulative percentage of variance		16.47	28.66	40.16	49.78	59.18	67.51
Cronbach alpha coefficient		0.76	0.74	0.70	0.63	0.59	0.58

5 Conclusion

A majority of factors - fifteen - were rated as critical (rating ¿ 4). Only one factor, ERP quality was not rated as critical in this study (rating ¡ 3.0). We found out those who rated ERP quality as poor were running tier 2 ERP systems which have not matured. Hence, the perceptions on challenges of Kenyan managers involved in ERP systems implementation projects are largely consistent with [9] findings that ERP projects faces additional challenges in developing countries related to economic, cultural and basic infrastructure issues. In fact, ERP implementations in Kenyan organisations face some unique challenges. This research has derived six composite challenges in ERP systems implementation in Kenya:

1. Integration and staff turnover issues
2. High cost further escalated by extensive customization
3. Poorchange management and failure to realize ERP benefits
4. Unreliability of vendors and poor quality of some ERP systems
5. Lack of skills by both users and consultants
6. Complexity of ERP system further compromising its security

The conclusions from the paper suggest that there are several fruitful avenues for future research. The empirical evidence revealed the need for improving the education system in order to help developing the required IT skills to particularly implement and use ERP software in Kenyan organisations. There is also imperative that the top management assume responsibility and drive change management throughout the implementation cycle.

There are a number of questions still to be determined. For example, although this paper establishes the relative importance of challenges faced by Kenyan organisations, it has not established the reasons. Future studies could look at differences by size of firms, by industry type, by number of locations, by number of customers etc.

The conflict, caused by the implementation of ERP systems in developing countries, between the globalization and localization of management work practices is worth of further investigations. However, this conflict is not a unique phenomenon to developing countries.

References

1. Zhang, M.K.O., Lee, L.: Critical success factors of enterprise resource planning systems implementation success in china. In: Proceedings of the 36th Annual Hawaii International Conference on System Sciences (2003)
2. Arif, M., Kulonda, D., Jones, J., Proctor, M.: Enterprise information systems: technology first or process first? Business Process Management Journal 11(1), 5–21 (2005)
3. Davenport, T.: Putting the enterprise into the enterprise system. Havard Business Review, 121–131 (July/August 1998)
4. Kumar, K., Hillegersberg, J.V.: Erp experiences and evolution: Introduction. Communications of the ACM 43(4), 22–26 (2000)

5. Allen, J.P.: Value conflicts in enterprise systems. Information Technology & People 18, 33–49 (2005)
6. Alshawi, S., Themistocleous, M., Almadani, R.: Integrating diverse erp systems: a case study. The Journal of Enterprise Information Management 17(6), 454–462 (2004)
7. O'Kane, J.F.: Erp implementations and cultural influences: a case study (2002)
8. Davison, R.: Cultural complications of erp valuable lessons learned from implementation experiences in parts of the world with different cultural heritages. Communication of the ACM 45(7), 109–111 (2002)
9. Huang, Z., Palvia, P.: Erp implementation issues in advanced and developing countries. Business Process Management Journal 7(3), 276–284 (2001)
10. Kremers, H., Dissel, M.: Erp system migrations. Communications of the ACM 43(4), 53–56 (2000)
11. Kumar, B., Maheshwari, V.: An investigation of critical management issues in erp implementation: emperical evidence from canadian organizations. Technovation 23(10), 793–807 (2003)
12. Chen, I.J.: Planning for erp systems: analysis and future trend. Business Process Management Journal 7(5), 374–386 (2001)
13. Gulla, T., Brasethvik, J.A.: A model-driven erp environment with search facilities. Data & Knowledge Engineering 42(3), 327–341 (2002)
14. Markus, L.M., Tanis, C., Fenema, P.C.V.: Multisite erp implementations. Communications of the ACM 43(4), 42–46 (2000)
15. Skok, W., Legge, M.: Evaluating enterprise resource planning (erp) systems using an interpretive approach. Knowledge and Process Management 9(7), 72–82 (2002)
16. Aladwani, A.M.: Change management strategies for successful erp implementation. 7(3), 266–275 (2001)
17. Al-Mashari, M., Al-Mudimigh, A.: Erp implementation: lessons from a case study. Information Technology & People 16(1), 21–23 (2003)
18. Johnston, S.: Erp: Payoffs and pitfalls, tech. rep., HBS Working Knowledge (2002)
19. Gosain, S.: Enterprise information systems as objects and carriers of institutional forces: The new iron cage? Journal of the Association for Information Systems 5(2), 151–182 (2004)
20. Al-Mashari, M., Zairi, M.: Information and business process equality: The case of sap r/3 implementation. Electronic Journal of Information Systems in Developing Countries 2(4), 1–15
21. Chatfield, A., Collins, C.: Introduction to Multi-Variate Analysis. Chapman & Hall, London (1992)
22. Kim, C., Mueller, J.: Factor Analysis: Statistical Methods and Practical Issues. Sage Publications, CA (1978)
23. Nunnally, J.: Psychometric Theory. Prentice-Hall, Englewood Cliffs (1987)

Web Data Integration System: Approach and Case Study

Abdolreza Hajmoosaei and Sameem Abdul-Kareem

Faculty of Computer Science and Information Technology
University of Malaya, PO Box 50603, Kuala lumpur, Malaysia
reza_moosaei@yahoo.com, sameem@um.edu.my

Abstract. There are a lot of valuable data on the web that organizations or users can use to improve their decision making process. It is therefore, very important and critical that this information be complete, precise and can be acquired on time. Most web sources provide data in semi-structured form on the internet. The extraction and combination of semi-structured data from different sources on the internet often fails because of syntactic and semantic differences. The access, retrieval and utilization of information from the different web data sources imposes a need for the data to be integrated. Integration of web data is a complex process because of the open, dynamic and heterogeneity nature of web data. The solution to this problem is a web data integration system. External information can be extracted from web sources and utilized for users through a web data integration system. In this paper, we first propose an approach and architecture for web data integration system and then develop a prototype of the proposed system for Malaysian universities.

Keywords: Web data source, Heterogeneity conflict, Web data integration.

1 Introduction

The web is the platform for information publishing; it is the biggest resource of information of any type. There are a lot of valuable data and business data on the web that organizations or users can use to improve their decision making process. It is therefore, very important and critical that this information be complete, precise and can be acquired on time [1]. It is also vital that such external information be systematically managed and utilized for users. The solution to the mentioned problem is a web data integration system. External information can be extracted from web sources and utilized for users through a web data integration system. The access, retrieval and utilization of information from the different data sources imposes a need for the data to be integrated. There are many types of heterogeneity and differences among web sources that make a combined effort to access data from different sources on the internet difficult and error-prone [2] [3].

In web data integration process we need to resolve heterogeneity conflicts between web data sources. There are different views about classification of Heterogeneity conflicts. The heterogeneity conflicts can be classified according to the following abstraction levels [2] [4]:

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 410–423, 2008.

- *Data Value Conflicts:* Data value conflicts are those conflicts that arise at the instance level. They are related to the representation or the interpretation of the data values. Examples of these conflicts are discrepancies of type, unit, precision and allowed values (e.g. "kg" and "gram" or "$" and "dollar").
- *Schema Conflicts:* Schema conflicts are due to different alternatives provided by one data model to develop schemas for the same reality. For example, what is modeled as an attribute in one relational schema may be modeled as an entity in another relational schema for the same application domain (e.g. "Author" as attribute for the entity "book" and "author" as an entity that has a relationship with "book"). Another example two sources may use different names to represent the same concept (e.g. "price" and "cost") , or the same name to represent different concepts , or two different ways, for conveying the same information(e.g. "data of birth" and "age").
- *Data Model Conflicts:* Data model conflicts occur when databases use different data models, e.g., one database designed according to the relational model, and another one object-oriented.

Conflicts in each level can be categorized into two categories:

- *Syntactic Conflicts:* Syntactic conflicts refer to discrepancies in the representation of data (e.g. "1/23" and "1.23" or "price=23$" and "price: 23$").
- *Semantic Conflicts:* Semantic conflicts refer to disagreement about the meaning, interpretation use of the same or related data (e.g. "staff" and "employee").

The major aim of our work is to give a solution for resolving heterogeneity conflicts mentioned above in a web data integration process. For this purpose we first recommend an approach and architecture for web data integration system and subsequently develop a prototype of our web data integration system in domain of Malaysian universities.

2 System Approach and Architecture

There are approaches for web data integration that have been proposed by researchers. Some of the major researches and approaches in web data integration domain are SIMS [5], COIN [6], MOMIS [7], KRAFT [8], OBSERVER [9]. For reconciliation of semantic conflicts between heterogeneous data sources, the above mentioned projects create one global or shared ontology by integrating or merging local schemas or ontologies. Subsequently, they perform semantic mapping between created global ontology and the local schemas or ontologies. In the web context the maintenance and updating of global or shared ontology is very time consuming and costly because many web data sources are involved and the number of involved web data sources change frequently; web designers and users are free to use their own terms and vocabulary and schemata which are subject to frequent changes. In our proposed approach we try to overcome this problem by using domain specific ontology and we resolve semantic problems between web sources through semantic mapping between the domain ontology and the local ontologies.

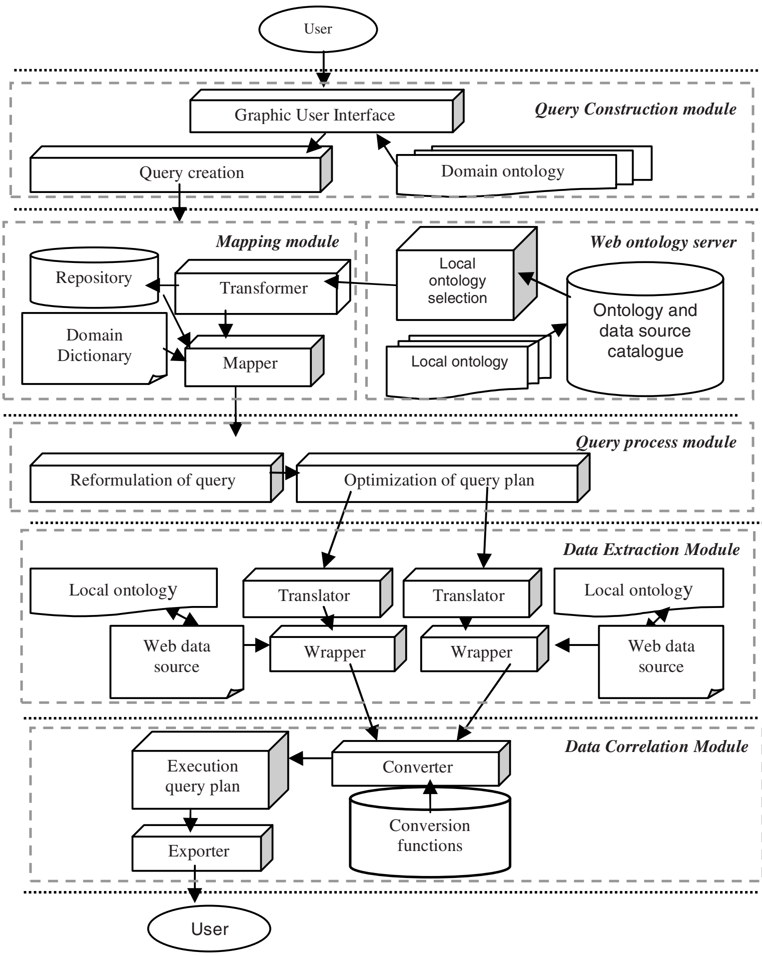


Fig. 1. Web Data Integration System

Our proposed system (figure1) uses domain specific ontologies for the creation of user queries. There is a domain specific ontology for each application that covers the semantic definition of terms which are required for user queries in a particular application domain. The domain ontologies are modeled in a uniform representation model. The user can browse the domain ontology and choose terms for his/her query; afterwards the system creates the user query. We assume each web source has an underlying pre-existing local ontology on the web and each local ontology is associated with one or more web sources. After the creation of the user query, the web ontology server chooses local ontologies related and relevant to user query domain and sends them to the mapping module. The first, local ontology is transformed to the system uniform representation model by transforming and mapping to corresponding terms from the local ontology and thence subsequently rewritten using the terms from the local ontology. This system uses ontologies to resolve semantic schema conflicts between web data sources. The proposed system resolves semantic schema heterogeneities between web source and user query through semantic mapping between the domain ontology and local ontology. The rewritten user query is sent to a query process module for reformulation and creation of optimized query plan from the user query. The gained sub queries from reformulation process are translated to the web sources query languages by translators and then the answer of sub-queries are extracted from related web sources by wrappers. Each wrapper knows the structure of the underlying web source and extracts the related data to its sub-query and presents the extracted data in a common format. Through wrappers, the system resolves data model heterogeneity conflicts between web data sources.

The extracted data obtained through wrappers are sent to the converter. The converter resolves data value heterogeneity conflicts between data values. The converter exploits conversion functions and rules for resolving heterogeneities between two data value (such as units, data types, value format conversion functions). Finally the query plan operations are executed and final answers are exported to user preference format by the exporter.

Our proposed web data integration system covers all abstraction levels of data heterogeneity conflicts between web data sources. The system applies:

- ontology as a solution for resolving schema heterogeneities;
- wrapper as solution for resolving data model heterogeneities;
- converter as solution for resolving data value heterogeneities;

The proposed web data integration system is scalable to any domain by adding related domain ontology to system. That means for using of the system in any application area we must develop and add a domain ontology (relevant to application domain area) to system. Our system implements a query based approach to information extraction and integration, from heterogeneous and distributed web data sources.The extraction and integration process in the proposed system consists of eleven major tasks as follows:

1. Creation of user query;
2. Determination of related local ontologies and their underlying web sources with query domain;
3. Transformation of related local ontologies to internal uniform representation model;
4. Semantic mapping between query terms and related local ontologies terms;
5. Rewriting of user query with corresponding terms from local ontologies;
6. Reformulation of query and creation of optimized query plan;
7. Translation of sub queries to web sources query languages;
8. Extraction of data;
9. Conversion of data values;
10. Execution of query plan operations;
11. Exporting of data;

3 Prototype of System

In the rest of the paper we discuss a prototype of our proposed web data integration system in a specific domain. We develop a system for universities in Malaysia so that users can access their required information relating to Malaysian universities. For example students can pose queries about programs, admission dates, fees or lecturers can access information about new publications, events and so on, in Malaysia universities.

3.1 Building University Ontology

Before the creation of the domain ontology we need to define one uniform representation model for the ontologies in our system. In order to compare and find similar terms between the domain and local ontology, the system needs to represent all ontologies in a uniform model. In our system we propose one uniform representation model for ontologies. This representation model is general and any ontology with any representation model can be transformed to this uniform representation model.

Definition1: T:=(C,A,R,V), each ontology element (term) is one of following entities:

- C: concept or instance of one concept
- A: attribute of one concept
- R: relationship between concepts
- V: value range of one relationship

For example student (concept), age (attribute), master student (instance of student, it is considered as sub-concept of student in our model), attend (relationship between student and class) and “<20” (value range of “max-credit-course” relationship) are some terms of university ontology.

Definition2: C:=(name, syn-set, A, key-A, key-R), each concept is defined with its name, set of its synonyms, attributes, its key attributes, and key relationships with other concepts. The key attributes and key relationships are subset of concept attributes and relationships. The key attributes and key relationships are specific properties of one concept that characterize the concept (e.g. teaching for lecturer). The key properties are specified only for domain ontology concepts during the creation of the domain ontology.

Definition3: A:=(name, syn-set), attribute is defined with a name and a set of synonyms.

Definition 4: R:=(name, syn-set, domain, range), each relationship is defined with a name, set of synonyms and the domain and range.

Definition 5: V:=(value), this feature is used for representing the range of one relationship that is a value. One value begins with one of these characters: “=”, “<”, “>” or “< >” and one string that show the value of its range.

Definition 6: $O:=(G, G^{'})$, each ontology is represented by two graphs.

Definition 7: $G:=(N,E)$, $N=<C>$, $E=<is\text{-}a>$, G is acyclic directed rooted graph that consists of nodes and edges. Each node is a concept (or instance of a concept). Each edge is *"is-a"* relation that shows sub-concept (subclass) relation between nodes. Indeed, *G* is a hierarchy concept model of ontology. Each node has one father and may have no, one or more child nodes. If one node has two fathers, the model resolves this problem with repeating child node for each one of its fathers.

Definition 8: $G^{'}:=(N,E^{'})$, $N=<C,V>$, $E^{'}=<R>$, $G^{'}$ is cyclic graph that consists nodes and edges. Each node is a concept (or instance of a concept) or one value. Each edge is relationship between two nodes that show the relationship between concepts. Indeed, G' is a concept relationship model of ontology.

The synonym sets are specified only for domain ontology terms in during the creation of the domain ontology. In our uniform representation model, all elements (concepts, attributes, relationships and values) are string (chain of characters). Our representation model of ontology is very general, so that our proposed approach which uses this formalization will work with any ontology representation languages. We need to transform the local ontology to the uniform representation model (in mapping module of system).

We build computer college domain ontology based on our uniform representation model. We follow the methodology below for building the domain ontology.

Phase 1: Domain and Scope clarification. Scope of our domain ontology is the current existing universities in Malaysia.

Phase 2: Concept (or instance), attribute and relationship extraction. For extraction of existing terms in the university we need to investigate and study conceptualization of this domain in detail. In our investigation we should answer several basic questions:

- Why are we using the ontology?
- For what types of questions should the ontology provide answers?

Phase 3: Key property determination. We specify key properties of each concept in domain ontology. These key properties show main meaning and semantic of each concept which characterizes it from others.

Phase 4: synonym determination. We exploit dictionaries and universities websites to specify synonyms of each existing term in the domain ontology.

Phase 5. Modeling: In this step we build concept hierarchy and relation concept graphs by an ontology-editing tool such as Protégé-2000, Ontolingua-1997 or Chimaera-2000. We use Protégé-2000 for our university domain ontology. Protégé-2000 was developed by Mark Musen's group at Stanford Medical Informatics.

Phase 6: Implementation. Finally we implement and store uniform representation models (hierarchy concept and relation concept graphs) of the university domain ontology in the SQL/SERVER DBMS.

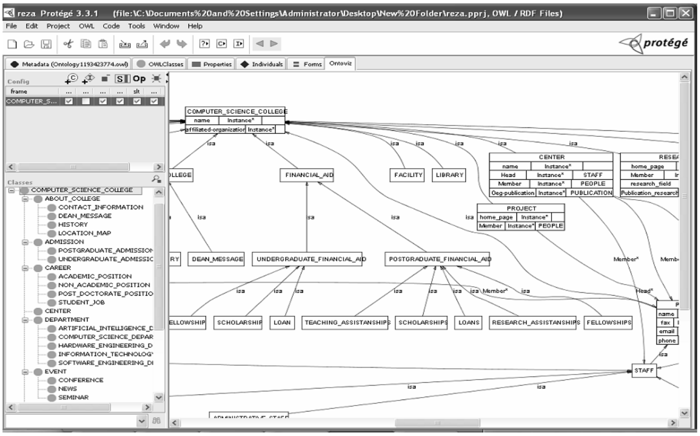


Fig. 2. Ontology editing tool

3.2 Building of GUI

Users interact with the system through a GUI (graphic user interface). The GUI must display the domain ontology to the user so that the user can find his/her query terms easily and quickly. After the user has found his/her query terms, the system creates the user query. The query construction module uses the following structure and syntax for the expression of the user query.

SELECT < *attributes names* >
FROM < *concept name* >
WHERE
{<*attribute name*$_1$: *values*>
<*attribute name*$_2$: *values*>
...}

In this query structure, the user can query the attributes of one concept from the domain ontology. The user can specify constraints and conditions over attributes of the query concept. Constraints and conditions are expressed after the "WHERE" clause in the query expression. We clarify the query syntax and structure with the following example:

Suppose that a user needs the names and emails of professors in university who are above 50 years of age and are female. The user traverses the terms in the university domain ontology and chooses his/her query terms. Afterwards, the system constructs an expression based on the user query as follows:

```
SELECT   name, Email
FROM      Professor
WHERE{ sex =Female , age >50 }
```

Note that the user can not pose complex queries. The user can split his/her complex query to simple sub-queries and subsequently submit them to the system. In our prototype, the GUI first displays super concepts of computer college domain ontology to the user (concepts in top level of hierarchy concept graph). The user chooses one of the super concepts. We call it, T_1. In next step the GUI shows three types of information to the user related to T_1 that are choices for the user:

- Sub-concepts (children) of T_1,
- Attributes of T_1,
- Relationships and their ranges which T_1 is the domain of those relationships (we call, relationship of T_1).

Figure 3 provides a partial illustration of this information about the concept of *student*.

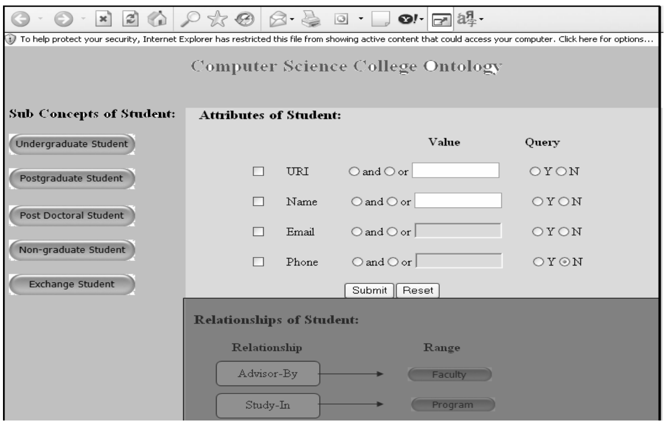


Fig. 3. Example of GUI

The user has the following choices for selection:

- If T_1 is a query concept, then:
  - user specify attributes of T1 which are in question and choose "Y" for query attributes and
  - If query possesses constraints then user selects attribute constraints and enters the constraint value of the attributes in the value fields. Afterwards, user submits form and the process of query construction is finished.
- If T_1 is not query concept then user repeats the following tasks until query concept is reached:
  - choose one of sub-concepts or
  - choose one of ranges (if range is concept element, no value element).

In this way, user finds his/her query terms and the query construction module creates the user query.

3.3 Developing of Ontology Server

The web ontology maintains a catalog of related ontologies and web sources, which contains information about local ontologies and the data sources underlying each local ontology. The ontology server is invoked by other modules for the following services:

- Selection of related local ontology with user query domain and transfer it to mapping module;
- Give information about existing data in related web sources to query process module;
- Give information about query language of related web sources to translators;

In this step, the first we list all universities in Malaysia and determine their ontologies. We should create ontology for those universities which have no ontology. All of Malaysian universities do not have ontology. We extract a concept hierarchy and concept relation to build the ontology for them from their web sites. We formalize extracted universities local ontologies in OWL (ontology web language) and maintain them in the ontology server. We used IIS as a server platform and implement the ontology server services by visual studio .NET framework.

3.4 Building Transformer

The universities local ontologies are formalized in OWL. We need to transform OWL-based local ontologies to a system uniform representation model. Our created transformer transforms the universities local ontologies to a system uniform representation model.

3.5 Mapping Algorithm

After the creation of user query by the query construction module, the query needs to be translated to relevant web sources query languages. For this translation first the user query terms must be semantically mapped to similar terms in the local ontologies of web sources. This semantic mapping resolves semantic schema conflicts between query terms and web source terms. The semantic mapping relates each query terms with its semantically similar term of related local ontology. The semantic mapping is partially performed between the domain ontology (related to query terms) and local ontology.

There are approaches for semantic mapping between ontologies that have been proposed by researchers. Some of the recent researches and approaches in ontology mapping domain are Chimaera [10], Anchor-PROMPT [11], QOM [12], Cupid [13], GLUE [14], SAT [15] and ASCO [16]. The approaches exploit available information from ontologies and map similar terms of two given ontologies to each other through mapping algorithms. Our mapping algorithm was motivated by some ideas of the above approaches.

Inputs of our mapping algorithm are: query terms, domain ontology and local ontology. There are three types of ontology term in user query: concept (*C*), attribute (*A*) and value (*V*). The purpose of the mapping algorithm is to discover semantically similar terms with query terms from the local ontology and then rewriting of queries with discovered similar terms. For calculation of similarity between two terms between user query and local ontology, the following function is used in mapping algorithm:

MF (Mapping Function): *$MF(T_1,T_2):=[0\ 1]$*; This function calculates similarity between two terms. Value rang of [0 1] indicates amount of similarity. MF performs two sub-functions for similarity calculation. First sub-function, normalizes two terms to their tokens. In this sub-function each term (concept, attribute) be:

- Tokenized: <Admission-fee> → <Admission, Fee>
- Lemmatized: <Fees> → <Fee>
- Eliminated: <subjectforprogram> → <Subject, Program>

For normalization, it exploits and uses one domain specific dictionary. This dictionary consists all existing terms in a specific domain including their synonym sets. The second sub-function compares tokens (string without space) of normalized terms with each other and calculates similarity between tokens. Finally, similarity between two terms is calculated from aggregation of token similarities. There are well-known metrics for calculating string similarity between two tokens such as Jaro-Winkler metric [17], Levenstein metric and Monger-Elkan [18]. We implement Jaro-Winkler metric in our mapping algorithm. The main Steps of our mapping algorithm are as follows:

First step: name and synonym matching between C and all local ontology concepts(C_L); MF is executed between *C(name)* (query concept in user query), all its synonyms names *C(syn-name)* with all university local ontology concepts (all *C_L(name)*).

```
for all C_L(name)<>null  do
{If MF(C(name),C_iL(name)) >= threshold  then
  Add (C, C_iL) to similarity-table;
 Else:  for all  C(syn-name) <>null  do
   {If MF(C(syn-name),C_iL(name)) >= threshold
     then Add (C,C_iL) to similarity-table; }}
```

The result of the above step is some similar pairs, which have similarity measure above the algorithm threshold. We call them, candidate mapping pairs.

Second step: father matching; in this step father and grandfather of *C* are compared with father and grandfather of each C_{iL} (C_{iL} in similar table). For each C_{iL} we calculate:

```
Father-matching(C,C_iL)←  [MF(father(C),  father(C_iL))  +
 MF(grandfather(C), grandfather(C_iL))]/2;
```

Third step: key property matching; Algorithm executes MF between key attributes and key relationships of *C* with all attributes and relationships of C_{iL}. (C_{iL} in similar table) Algorithm saves number of matching attributes and relationships (properties) between *C* and each C_{iL}.

```
while  C_iL (name)<>null   do
    { for all  A,R of C  do
```

```
If each MF(C(key-property-name & key-property-
 syn-names), CiL(A & R)) >= threshold then
 CiL(similar-property) +1;}
```

Fourth step: aggregation; In this step the algorithm finds the most similar corresponding concept with *C* among all C_{iL} in the similar table. For this purpose, algorithm aggregates results of second and third steps for each C_{iL}. The C_{iL} which has highest weight above the algorithm threshold is final mapping concept of *C*. We call it C_{1L}.

```
For all CiL do
{ Weight(CiL) ← father-matching(CiL) + CiL(similar-
 property); }
C1L ← CiL [Max weight (CiL)];
concept-mapping-table ← (C1, C1L);
```

Fifth step: Attribute mapping between C and C_{1L}; After finding similar concept of *C*, we must find similar attributes for query attributes and constraint attributes expressed in user query from local ontology. We should notice, algorithm just execute MF between *C-atrribute-name* (in user query), all *C-att-synset-names* with *attributes-names* and *relationships-names* of its mapping pair (C_{1L} in mapping table). Algorithm chooses maximum MF that is above threshold and stores similar-attribute pairs in *C-att-mapping table* (such as: <C-A_1,C_{1L}-A_{1L}> , <C-A_2,C_{1L}-A_{2L}>.....).

```
For all  C(A-name) in user query  do
  {  If    MF(C(A-name),C1L(A-name  or  R-name))  >=
        threshold  then
        Add  (C(A-name),  C1L(A-name  or  R-name))  to
            att-mapping-table;
     Else:  while  A-syn-name <>null  do
        {  If   MF(C(A-syn-name),C1L(A-name  or  R-name))
           >= threshold  then
           Add  (C(A-name),  C1L(A-name  or  R-name))  to
           att-mapping-table;   } }
```

If algorithm does not find similar terms of query concept and query attributes from the local ontology, mapping is not executed between the user query terms and the local ontology terms and the local ontology has failed.

3.6 Developing of Query Process Module

This module consists of the following tasks:

1. Reformulation of user query. After the mapping process, the user query that has been rewritten into terms of one or more local ontologies is reformulated into one or more sub queries. The query process module must first reformulate the query into sub queries that refers directly to the schemas in the web sources. In order for the system to do this, it needs to have a set of source descriptions. This information exists in the web ontology server.
2. Creation and optimization of the query plan. After the minimal set of data sources were selected for a given query and the query was reformulated to some sub queries a key problem is to find the optimal query execution plan for these sub

queries. In particular, the plan specifies the order in which to perform the different operations between the sub queries (join, selection and projection) and scheduling of different operators.

In our prototype we submit whole rewritten user query to each selected university web source. Therefore we do not need to reformulate the user query and query plan.

3.7 Building of Translators

After the query was rewritten with local ontology terms, it is translated to the local web sources query language. For each term of the web source there exists some mapping information in the repository of the web ontology server module that links the local ontology elements with the underlying data elements of the web source. This information is used by the translator to translate the user query (expressed in terms of the local ontology) into different queries to the underlying web sources query language.

In our prototype, the university local ontologies terms use the same name with their corresponding terms in the university web sites. In our prototype, the rewritten user query with universities local ontologies terms is sent to the related wrapper and the user query is parsed to each university web source query language by wrappers.

3.8 Building Wrappers

In this step, data is extracted from the pre-selected web sources by wrappers. Wrappers are used for retrieving data from web sources. A wrapper is a module which understands a specific data organization. It knows how to retrieve data from the underlying repository and hide the specific data organization to the rest of the information system. There are three freeware wrapper tools which have been developed by various researchers:

1. W4F (http://cheops.cis.upenn.edu/W4F/)
2. XWRAP (http://www.cc.gatech.edu/projects/disl/XWRAPElite/)
3. DEByE (http://www.lbd.dcc.ufmg.br/~debye/)

We customize one of above wrappers for our prototype.

3.9 Building of Converter

After extraction of answers by wrapper from web sources, they are sent to the converter. The converter resolves existing data value heterogeneities between extracted data values. The popular data value conflicts consist of discrepancies in:

- Type (e.g. integer and char)
- Unit (e.g. "kg" and "gram")
- Value format such (e.g. date format, time format).
- Value sign (e.g. "$" and "dollar")

For resolving these kinds of conflicts, the converter uses some conversion functions. These functions should cover all existing value data heterogeneities in the domain of the system. In our prototype domain, we define some conversion function for resolving conflicts between data values such as name, URI, date and fee. We define standard format for these values and convert extracted data values which are different with our standard format by conversion functions.

3.10 Building Exporter

Finally the data is ready to be presented to the user. In this step, the data is exported to the user representation environment in our prototype. We present data in the HTML page to the user.

4 Conclusion

In this paper, we first proposed an approach and architecture for a web data integration system. The proposed system resolves data heterogeneities between web data sources. Afterwards, we discussed the development of prototype and a case study for our proposed system. Our prototype is in the domain of the Malaysian universities. The existence of the proposed web data integration system is critical for any organization to access valuable external information on the web. Our proposed system extracts this valuable information from web sources, integrates them to each other and presents to the user.

References

1. Heflin, J., Hendler, J.: 2000: Semantic interoperability on the web. In Extreme Markup Languages (2000), http://www.cs.umd.edu/projects/plus/SHOE/pubs/extreme2000.pdf
2. Kashyap, V., Sheth, A.: Semantic heterogeneity in global information systems: The role of metadata, context and ontologies. In: Papazoglou, M.P., Schlageter, G. (eds.) Cooperative Information Systems: Current Trends and Directions, pp. 139–178. Academic Press Ltd, London (1998)
3. Fensel, D.: Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. Springer, Heidelberg (2001)
4. Ram, S., Park, J.: Semantic Conflict Resolution Ontology (SCROL): An Ontology for Detecting and Resolving Data and Schema-Level Semantic Conflicts. IEEE Transactions on Knowledge and Data Engineering 16(2), 189–202 (2004)
5. Arens, Y., Ciiee, Y., Knoblock, A.: SIMS: Integrating data from multiple information sources. In: Information science institute, University of Southern California, U.S.A (1992)
6. Goh, C.H., Bressan, S., Madnick, S., Siegel., M.: Context interchange New features and formalisms for the intelligent integration of information. ACM Transaction on Information Systems 17(3), 270–290 (1999)
7. Beneventano, D., Bergamaschi, S., Guerra, F., Vincini., M.: The MOMIS approach to information integration. In: ICEIS 2001, Proceedings of the 3rd International Conference on Enterprise Information Systems, Portugal (2001)
8. Visser, P.R., Jones, D.M., Beer, M., Bench-Capon, T., Diaz, B., Shave, M.: Resolving ontological heterogeneity in the KRAFT project. In: Bench-Capon, T.J.M., Soda, G., Tjoa, A.M. (eds.) DEXA 1999. LNCS, vol. 1677, pp. 668–677. Springer, Heidelberg (1999)

9. Mena, E., Kashyap, V.: OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies (1996)
10. McGuiness, D.L., Fikes, R., Rice, J., Wilder, S.: The chimaera ontology environment. In: McGuiness, D.L., Fikes, R., Rice, J., Wilder, S. (eds.) Seventh National Conference on Artificial Intelligence (AAAI-2000) (2000)
11. Noy, N.F., Musen, M.A.: Anchor-PROMPT: Using Non-Local Context for Semantic Matching. In: Workshop on Ontologies and Information Sharing. IJCAI, Seattle, WA (2001)
12. Ehrig, M., Staab, S.: Efficiency of Ontology Mapping Approaches. In: Institute AIFB, University of Karlsruhe (2001)
13. Madhavan, J., Bernstein, P.A., Rahm, E.: Generic Schema Matching with Cupid. Proc. Of the 27th Conference on Very Large Databases (2001)
14. Doan, A., Madhavan, J., Domingos, P., Halevy, A.: 2002: Learning to Map between Ontologies on the Semantic Web. In: The Eleventh International World Wide Web Conference (WWW 2002), Hawaii, USA (2002)
15. Giunchiglia, F., Shvaiko, P.: Semantic Matching. CEUR-WS 71 (2003)
16. Thanh Le, B., Dieng-Kuntz, R., Gandon, F.: On Ontology Matching Problems: for building a corporate Semantic Web in a multi-communities organization. In: Institute National and Research in Informatic, Sophia Antipolis, France (2004)
17. Winkler, W.E.: The state of record linkage and current research problems. Statistics of Income Division, Internal Revenue Service Publication R99/04 (1999)
18. Cohen, W., Ravikumar, P., Fienberg, S.: A Comparison of String Distance Metrics for Name Matching Tasks. In: IJCAI 2003, workshop on Information Integration on the Web. (2003)

Integration of Data and Rules in Inference with Queries Method

Andrzej Macioł

Faculty of Management, AGH University of Science and Technology,
ul. Gramatyka 10, 30-067 Kraków, Poland
amaciol@zarz.agh.edu.pl

Abstract. The paper presents the effects obtained so far by the author in integration of data and rules in Expert Systems and designing complex information systems. The Inference with Queries (IwQ) method was used. This solution is based on using SQL mechanisms with data stored in a relational database as a knowledge base. Our solution joins the concept of frame-based knowledge representation, replaced partially with a relational database model, with inference possibilities given by procedural languages. Results of the tests show that the idea of joining the possibilities of rule systems and SQL is advisable. The prospects connected with using IwQ method for designing complex information system and their simulation seem to be highly interesting.

Keywords: rule-based systems, SQL, system modeling, business information systems.

1 Introduction

Integration of data and decisions is the main problem in designing and implementation of modern information systems. This refers not only to solutions focusing on supporting or substituting decision-making processes (Expert Systems, Control Systems and others) but also to those solutions, of which the main task is to support communication processes in organizations (ERP, SCM, Workflows, etc). In the first case, persistent, but at the same time flexible recording of facts and rules is a problem. The problem is solved by either using database mechanisms or ontology management tools.

In the case of designing and implementation of systems concentrated on processing data in organizations we now have a wide variety of solution modeling and generating tools, among them BPM and UML, which have become very popular in recent years. Unfortunately, they have limited possibilities of mapping complex decisions and decision processes. It is widely thought that integration of different modeling ideas from AI techniques used in expert systems, via database management and ontology tools, to data and process modeling methods and languages (BPM, UML) is necessary.

One of such solutions is our Inference with Queries method (IwQ)[1]. This solution is based on using SQL mechanisms with data stored in a relational

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 424–435, 2008.

database as a knowledge base. Our solution joins the concept of frame-based knowledge representation, replaced partially with a relational database model, with inference possibilities given by procedural languages (Transact-SQL in current version of the solution) and is closethe rule-based languages idea [2].

The paper presents the effects obtained so far in integration of data and rules in Expert Systems and designing complex information systems.

2 Foundations for Inference with Queries Method

The source of IwQ method came from considering the possibility of developing tools for modeling business rules management systems (BRM). It turned out that decision problems depicted by business rules usually belong to the class of unstructured problems. The knowledge about such problems cannot be represented in a procedural manner. As a result, only those methods of knowledge representation, which allow building declarative model of decisions, can be used. One of these methods is a frame-based approach to knowledge formulation [3]. Our solution is very closethis idea but instead of specific relations between atomic data implemented by frames and hierarchy facets ("A-kind-of"), as well procedural facets (e.g. "If-needed", "If-created",. . .) we use mechanisms of relational database (relationships, triggers, stored procedures etc).

Our model of the rule-based system use an extended form of the rules including both control statement and dynamic operations. Generic form of a rule can be presented as follows:

$$
\begin{array}{l}
rule(i) : (A_1\ r\ d_1) \wedge (A_2\ r\ d_2) \wedge ... \wedge (A_n\ r\ d_n) \\
\ \rightarrow \\
\quad set(B_1 = b_1, B_2 = b_2, ..., B_b = b_b) \\
\quad H_1 = h_1, H_2 = h_2, ..., H_h = h_h \\
\quad next(j) \\
\ else \\
\quad set(C_1 = c_1, C_2 = c_2, ..., C_b = c_b) \\
\quad G_1 = g_1, G_2 = g2, ..., G_h = g_h \\
\quad else(k)
\end{array}
$$

where $(A_1\ r\ d_1) \wedge (A_2\ r\ d_2) \wedge ... \wedge (A_n\ r\ d_n)$ is the regular precondition formula (A_i is the value of object's attribute, r is a relational symbol, i.e. =, > and d_i is a varable), $B_1 = b_1, B_2 = b_2, ..., B_b = b_b$ is the specification of the facts to be changed in knowledge base after successful execution of the rule, $H_1 = h_1, H_2 = h_2, ..., H_h = h_h$ is the specification of conclusions forming a direct output of the rule (e.g. decisions or queries to be displayed on the terminal or control actions to be executed) in case when it is successful executed, $C_1 = c_1, C_2 = c_2, ..., C_b = c_b$ is the specification of the facts to be changed in knowledge base in case of failure, $G_1 = g_1, G_2 = g2, ..., G_h = g_h$ is the specification of conclusions in case of failure and $next(j)$, $else(k)$ are the specifications of control; the next(j) part specifies which rule should be examined immediately after successful execution of rule i and $else(k)$ part specifies which rule should be tried in case of failure.

In our model objects, attributes and variables can be selected from any relational structure of RDBS (relation or join of relations). We use one object class to store all the necessary data. This object is called facet. The facet can have a value taken from a relational structure. Thanks to the utilization of extended selection formula, knowledge definition process gets simplified.

Our solution isn't a classical computer language. The process of model developing and debugging is realized with help of appropriate editor. We can describe IwQ syntax as the structure of entities representing both facets and rules.

The rules are stored in folowing entities:

$Rules = \{AgentType, RuleNumber, TrueRuleNumber, FalseRuleNumber\}$

This rule contains the agents' type identifier (only in multi-agent version of the system), rule number and the control part of the rule ($next(i)$ as ***TrueRuleNumber*** and $else(i)$ as ***FalseRuleNumber***).

$LeftHand = \{RuleNumber, FacetsNameLH, FacetsNameRH,$
$LHSource, Operator, RHSource\}$

This rules rows contain the full description of preconditions. For each rule we can define an unlimited set of preconditions containing names of left hand condition facets, right hand of condition facets and the relational symbol (***Operator***). Because a facet returns two values: the value in specified column and the index of row, columns ***LHSource*** and ***RHSource*** specify which value will be taken for examining.

$RightHand = \{RHIndex, RuleNumber, TrueOrFalse, Seq, ActionType\}$

The ***RightHand*** entity plays a control rule for actions executed after examining the left hand of the rule and specifies which action of ***ActionType*** will be executed in the sequence specified by ***Seq***.

We have four types of actions: events, procedures, operations changing the variables and/or facts and a special kind of procedures for a dialog with the users.

Each of action is depicted by one entity:

$RHEvent = \{RHIndex, AgentType, StartRuleNumber\}$

This action raises an event for an agent (***AgentType***) and indicates which rule (***StartRuleNumber***) should be fired first when the system handles this event.

$RHProcedure = \{RHIndex, ProcedureName, ProcType, ProcAdress\}$

Procedures are identified by name (***ProcedureName***) and type (***ProcType***), e.g. SQL procedure, the objects method got from an assembly etc. In column ***ProcAddres*** we can add an information necessary for finding the source of procedure, e.g. the assembly name. For each procedure we can specify the set of needed arguments in child entity ***ProceduresParameters***.

$ProceduresParameters = \{RHIndex, ParameterName\}$

The instruction how to change facts and/or variables values is specified in entity:

$RHOperations = \{RHIndex, InputFacet, InputSource, OutputFacet,$
$OutputDestination, Operation, OperationArgument\}$

Each row of this table includes names of input and output facets (***InputFacet***, ***OutputFacet***), information which output of the facet is to be taken into an operation - value or collection index (***InputSource***, ***OutputDest***) and a name of the predefined two operation arguments (***Operation***) with its argument (***OperationArgument***) being a facet's name. The last two data are necessary when a new value of output variable or fact is a result of a simple arithmetic operation or mathematical function.

A special kind of procedures allows to perform the dialog between the user and the inference engine. The System provides a set of forms for different methods of dialog, e.g. an atomic value input, selecting from the list, selecting multiple value by checkbox, showing an atomic value or a list of values, etc.

$RHAsk = \{RHIndex, InputFacet, OutputFacet, InfoTextFacet,$
$ConstText, AskType\}$

In this facet we can specify a name of the facet for each rule to be determined by a user (***InputFacet***), a name of the facet which has to be displayed on the terminal (***OutputFacet***), a name of the facet containing the value to be shown on the terminal (***InfoTextFacet***), the constant text to be shown on the terminal (***ConstText***) and a type of the form (***AskType***).

The central entity containing the information about all attributes, values, constants and facts in Knowledge Base is entity:

$Facets{=}\{AgentId, FacetsName, TableName, SelectedColumn, ValueColumn,$
$CollectionIndex, FunctionName\}$

Each object in knowledge base is identified by the unique name and - in multi-agent version of the system - by agent identifier. In case of attributes bounded with RDBS tables, the table name (***TableName***), a name of the column to return from a table (***SelectColumn***), perhaps a name of the column containing the auxiliary value[1] returning from a table (***ValueColumn***), the index of the collection of returned rows (***CollectionIndex***) and the name of the function used in a select expression, (***FunctionName*** e.g. SUM, MAX, etc) are specified. In case ***TableName*** column contains null value the system identifies variable, constant or fact. The ***FacetsValue*** contains the current value of the facet. For each facet bounded with RDBS tables we can specify an unlimited set of search conditions contained in entity:

$FacetsWhere = \{FacetsName, ParametersName, Operator, ColumnName\}$

The condition is described by the column name in table bounded with parent facet (***ColumnName***), the relational symbol (***Operator***) and the name of the facet which is the right hand predicate of search condition (***ParametersName***).

Currently, several versions of tools implementing IwQ idea were elaborated and tested. In every case, Microsoft SQLServer and T-SQL language were used

[1] This value is needed in case another value is shown on display and another represents the choice of a user, e.g. in combo boxes.

as the implementation environment. Sophisticated interfaces (partially based on GUI) have been developed for all cases.

3 Related Works

The enormous gap between business rules and their actual implementation in business software has long been identified as a major obstacle for integration of management information systems. The effort to integrate complex business rules with process and workflows models are in different measures taken in the research related to business process modelling (BPM), business rules management (BRM) and - recently - web semantics.

Among the BPM tools two important standardization efforts include the development of the Business Process Modeling Notation (BPMN) and the Unified Modelling Language (UML) version 2.0 [4]. The BPMN was created by the Business Process Management Initiative (BPMI) organization [5]. Extended BPMN models and UML Activity Diagrams models enable to record sufficient amount of information necessary for simulating system behaviour. Unfortunately, in most cases the type of modelling does not allow precise depicting of complex inference rules . The Business Rule Approach is intended to assert business structure, or to control or influence the behaviour of the business [6]. Most of the work done in this area focuses on identifying and documenting business rules [7], business rules lifecycle (acquisition, deployment and evolution) [8] and enterprise knowledge relationship to a software system [9]. From the IS design perspective the necessity of extracting system behaviour rules from application code gains higher significance. Such approach is characteristic for programming languages according to the programming with knowledge or rule-oriented programming concepts [10]. Business rules, as part of requirements gathering and systems analysis have not been ignored by structured analysis, information engineering or object-oriented analysis approaches [11]. A special importance of business rules specifying in information systems design attaches agile methods of modelling [12]. Many researchers have suggested approaches or ideas to integrate AI and databases. From the one side there are works concerned on intelligent databases as Datalog [13]). From the other hand, there are investigations to couple expert systems with relational or object databases [14]. Today's available shell expert systems (e.g. CLIPS, Jess) allow to retrieve and store facts, necessary for inference, in a database. However we don't know any ES software, that uses the possibilities of SQL in inference process. It appears that nearest to our idea is a concept of prefetching and caching decisions by SQL queries presented in [15] but in this case too, the queries do not replaced the inference procedures.

In last few years many researches developed and tested web-based expert systems. These systems use usually XML-based techniques to store the knowledgebase. In contemporary e-services the important role plays intelligent agent armed with knowledge and inference possibilities. Representation of these problems is shown in [16]. The research presented above gives examples and proves that the main problem of contemporary web-based intelligent system used as modelling tools or active agents is integration of ontology represented usually in RDF or OWL form with querying mechanism (e.g. SPARQL) and inference engines (Jess, CLIPS).

In comparison to other solutions, IwQ method is characterized by the following features:

– rules are recorded in the relational data model, which allows the user editing the rules freely, without any limitations of preconditions or conclusions,
– the user can use inference mechanisms in SQL to simplify their own rules freely,
– information necessary for inference (facets) can come from any external source, provided that their conversion to the form accepted by SQL is possible,
– inferring is driven by rules, which limits the possibility of new knowledge generation; on the other hand, it is possible to use this method as a procedural SQL extension,
– a set of rules connected with control addresses and starting a wide range of predefined and external procedures on the right hand is the core of a universal tool for decision and process modeling.

4 Application of IwQ Method in Expert Systems

The application of IwQ as a shell tool for Expert Systems allows to decrease a number of rules, which can be replaced with SQL queries. Moreover, it is possible to edit rules, as well as information sources (facts and variables).

Let us consider an example of a rule-based reasoning system for credit ability of bank customers evaluation. The declaration of one of these rules is as follows:

Rule: 1
if the credit facility is greater then 130% of credit **and** the profit is high and stable **and** all credits are repaid **then** financial state is very good
Rule: 2
if the credit facility is equal to credit **and** the profit is high and stable **and** all credits are repaid **then** financial state is very good
Rule: 3
if the credit facility is greater then 130% of credit **and** the profit is moderate **and** all credits are repaid **then** financial state is very good
Rule: 4
if the credit facility is greater then 130% of credit **and** ability to credit maintenance is good **and** the profit is high and stable **and** all credits are on time repaying **then** financial state is very good
Rule: 5
if the credit facility is greater then 130% of credit **and** ability to credit maintenance is good **and** the profit is moderate **and** all credits are on time repaying **then** financial state is very good

Rule: 6
if the credit facility is greater then 130% of credit **and** ability to credit maintenance is moderate **and** the profit is high and stable **and** all credits are on time repaying **then** financial state is very good

...

To establishing the financial state of the customer 53 rules are needed. Thanks our method after grouping of selected parameters in sets represented by proper joins of relations we can replace some rules by SQL queries. An adequate algorithm decides about grouping individual variables into joins. This algorithm considers minimization of the number of data necessary to obtain satisfactory answers in the inference process as the objective function. For example we can join the *profit level* with *credit facility* creating the perspective:

$ProfitFacility = \{Profit, FacilityFrom, FacilityUntil, HypothesisVal\}$
with following rows:

Profit	**FacilityFrom**	**FacilityUntil**	**HypothesisValue**
high and stable	130%	9999%	PF1
moderate	130%	9999%	PF1
high and stable	99%	101%	PF1
...			

and **credit history** with *ability to credit maintenance*:
$HistoryMaintenance = \{CreditHistory, MaintenanceAbility, HypothesisVal\}$
with following rows:

CreditHistory	**MaintenanceAbility**	**HypothesisValue**
on time repaying	good	CHMA1
on time repaying	moderate	CHMA1
...		

Thanks to them, that some of preconditions are included in selecting formulae we need only two rules for examining if the financial state of the customer is very good. The declaration of these rules is as follows:

$rule(1) : CreditHistory.v = CAllRepaid.v \wedge$
$\wedge\ SProfitFacility.v = CPV1.v$
$\rightarrow$
$set(FinancialState.v = CVeryGood)$
$next(3)$
$else$
$else(2)$

$rule(2) : SProfitFacility.v = CPV1.v \wedge$
$\wedge\ SHistoryMaintenance.v = CHMA1.v$

$\rightarrow$

$set(FinancialState.v = CVeryGood)$
$next(3)$
$else$
$else(3)$

where $CreditHistory.v$ is the value of the unbounded facet selected by the user by system interface, $CAllRepaid.v$, $CPV1.v$, $CVeryGood$ and $CHMA1.v$ are the constants, $SProfitFacility.v$ and $SHistoryMaintenance.v$ are the values of the facets bounded with proper queries, $FinancialState.v$ is the value of unbounded facet established by the inference engine.

Before the rules can be examined, the facets in the left hand side of the rules are converted into SQL query. In our case there are the following queries:

$SELECT\ HypothesisValue\ AS\ SProfitFacility.v$
$FROM\ ProfitFacility$
$WHERE(Profit = @Profit.v)$
$AND(@Facility\ BETWEEN\ FacilityFrom\ AND\ FacilityUntil)$

$SELECT\ HypothesisValue\ AS\ SHistoryMaintenance.v$
$FROM\ HistoryMaintenance$
$WHERE(CreditHistory = @CreditHistory.v)$
$AND(MaintenanceAbility = @MaintenanceAbility.v)$

Before the queries are executed the variables signed by @ symbol are converted into proper strings.

On the same manner all rules in the knowledge base can be simplified. At last instead of 53 rules we need 10 rules and 9 queries to solve presented problem.

5 Modeling Complex Information Systems

The main aim for creating IwQ method is modeling complex information systems. The method is useful for these complex problems, which, on one hand deal with the effectiveness of collecting and making information in workflows accessible, and on the other they require solving decision problems. Examples of these problems are manufacturing plan optimization, job shop scheduling and others, which are optimization problems characteristic for operational research, but, at the same time they must be tested in connection with mastering workflows, business process management and IT.

Designing Supply Chain Management System is just one of those problems. The procedure of realizing purchases in a large multinational production company can be an example. The tool made in accordance with IwQ method enables to construct and test many alternatives of the designed or improved system, taking into consideration both different decision rules or process and decision-making flowcharts.

Rules used to build a model can be divided into two groups. The first group represents business rules, the second one controls simulation process. Exemplary rules describing decisions about supply sources have the following formula:

$rule(20) : CurrentSourceIndex.v <= SourceIndexCount.v$
'end of supply sources list examining
$\rightarrow$
$set(PriceDiff.v = Substr(SSourcePrice.v, BestPrice.v))$
$set(PriceDiff.v = Divide(PriceDiff.v, BestPrice.v))$
$set(AbsPriceDiff.v = Abs(PriceDiff.v))$
$set(DeliveryTermsDiff.v =$
$Substr(SDeliveryTerms.v, BestDeliveryTerms.v))$
$set(DeliveryTermsDiff.v =$
$Divide(DeliveryTermsDiff.v, BestDeliveryTerms.v))$
$set(RecommendationDiff.v =$
$Substr(BestRecommendation.v, SRecommendation.v))$
$set(RecommendationDiff.v =$
$Divide(RecommendationDiff.v, BestRecommendation.v))$
$set(DeliveryTimeDiff.v =$
$Substr(BestDeliveryTime.v, SDeliveryTime.v))$
$set(DeliveryTimeDiff.v =$
$Divide(DeliveryTimeDiff.v, BestDeliveryTime.v))$
'the relative differences between current source's
'and the best one's parameters
$next(25)$
else
$H_2 = SelectSource(BestSourceIndex.v)$
'the procedure indicates the best supplier
$else(105)$

$rule(25) : PriceDiff.v < Zero.v$
$AND\ AbsPriceDiff.v >= DistinctPriceDiif.v$
'the price of current supplier is distinct less then the best
$\rightarrow$
$set(ChangeTheBest.v = True.v)$
'the current and the best supplier have to be swapped
$next(100)$
else
$set(ChangeTheBest.v = False.v)$
$else(30)$

$rule(30) : PriceDiff.v > Zero.v$
$AND\ PriceDiff.v < DistinctPriceDiif.v$
$AND\ PriceDiff.v > MediumPriceDiif$
$AND\ DeliveryTimeDiff.v > DistinctDeliveryTimeDiif$
'the price of current supplier is less then the best

'and the delivery time is distinct shorter then the best
$\rightarrow$
$set(ChangeTheBest.v = True.v)$
$next(100)$
$else$
$set(ChangeTheBest.v = False.v)$
$else(35)$

$rule(35) : AbsPriceDiff.v < EqualPriceDiif$
$AND\ DeliveryTimeDiff.v > DistinctDeliveryTimeDiif$
$AND\ DeliveryTermsDiff > DistinctDeliveryTermsDiff$
$AND\ RecommendationDiff.v > DistinctRecommendationDiff.v$
'the price of current supplier is quiet equal to the best
'and the delivery time is distinct shorter then the best
'and the delivery terms are distinct better then the best
'and the supplier's recommendations are distinct better then the best
$\rightarrow$
$set(ChangeTheBest.v = True.v)$
$next(100)$
$else$
$set(ChangeTheBest.v = False.v)$
$else(40)$

...
$rule(100) : ChangeTheBest.v = True.v$
'this rule swaps or not, best and current supplier
$\rightarrow$
$set(BestSourceIndex.v = CurrentSourceIndex.v)$
$set(BestPrice.v = SSourcePrice.v)$
$set(BestDeliveryTerms = SDeliveryTerms.v)$
$set(BestRecommendation.v = SRecommendation.v)$
$set(BestDeliveryTime.v = SDeliveryTime.v)$
$set(CurrentSourceIndex.v = Add(CurrentSourceIndex.v, One.v))$
$next(20)$
$else$
$set(CurrentSourceIndex.v = Add(CurrentSourceIndex.v, One.v))$
$else(20)$
'in both cases engine take the next supplier and return to the first rule

Presented above schema shows an bubble sort algorithm where more then one condition is used as ordering criteria. The constants depicting ***distinct*** and ***medium*** differences between conditions can be modified during simulation experiments.

Presented model can be used for simulation tests in business analysis. It could be also used as a framework for future implementation, based on executable modules and defined business rules, prepared by a user.

6 Conclusions

Results of the tests show that the idea of joining the possibilities of rule systems and SQL is advisable. The prospects connected with using IwQ method for designing complex information system and their simulation seem to be highly interesting. Our solution can't be compared directly with any available tools or languages. The closest ideas are realized in web based intelligent systems based on ontological representation of the knowledge (RWD, OWL) and first order predicate logic querying and reasoning tools (SPARQL, Jess). This solution requires XML-based facts and knowledge representation. Unfortunately today most of management information systems work on the base of relational databases. The conversion of database schema to ontology schema is possible but in our meaning this step is a not acceptable in problem solving. We foresee that in nearest future the business rules modelling systems based on relational databases and SQL will be the best solution for management information systems development. The advantages of this solutions are:

- combination of BPMN models simplicity, which was achieved thanks to maximum limiting the number of typical activities, with the possibilities of describing complex business rules in the manner characteristic for expert systems,
- simple logic structure of the model thanks to using facets as SQL queries,
- full representation of real environment, where most business processes take place, thanks to integration with SQL databases,
- easy on-line cooperation with ERP systems and user's own applications thanks to information exchange in SQL databases standard and easy use of user's own libraries,
- easy model conversion to any XML-based form, possible thanks to using relational schema for representation of the system and its behavior.

Partially replacement inference rules by SQL queries can effect workload time of database server. We tried to establish how important that problem is. Employing the problem example presented in section 4 we have tested how much time is required to solve the problem using, typical reasoning system made in algorithmic language. Next we have compared that to the case when most rules are replaced by SQL queries (IwQ method). Our tests show that overall execution time in the case when IwQ method was used can be longer till 14% with small set of test examples (less then 1000). The bigger examining set is, the more effective IwQ method appears. When the analyzed set size is greater then 3000 the execution time of IwQ solution is shorter than typical rule-based inference engine.

Further tests connected with the tool itself will now concentrate on improving and expanding its functionality with simultaneous preserving general assumptions about its operating rules.

References

1. Macioł, A.: An application of rule-based tool in attributive logic for business rules modeling. Expert Systems with Applications, Expert Syst. Appl. 34, 1825–1836 (2008)
2. Liu, M.: An Overview of Rule-based Object Language. J. Intell. Inf. Syst 10, 5–29 (1998)
3. Minsky, M.: A framework for representing knowledge. In: Winston, P. (ed.) The Psychology of Computer Vision., McGrow-Hill, New York (1977)
4. Hall, C., Harmon, P.: The 2006 Enterprise Architecture, Process Modeling and Simulation Tools Report http://www.bptrends.com
5. White, S.A.: Introduction to BPMN http://www.bpmn.org
6. Ross, R.G.: Principles of the Business Rules Approach. Addison-Wesley, Canada (2003)
7. Rai, V.K., Anantaram, C.: Structuring business rules interactions. Electron. Commer. R. A. 3, 53–74 (2004)
8. Rosca, D., Greenspan, D., Wild, C.: Enterprise Modeling and Decision-Support for Automating the Business Rules Lifecycle. Autom. Softw. Eng. 9, 361–404 (2002)
9. Wan-Kadir, W.M.N., Loucopoulos, P.: Relating evolving business rules to software design. J. Syst. Architect 50, 367–382 (2004)
10. Fowler, M.: Refactoring: Improving the Design of Existing Code. Addison-Wesley, Canada (2000)
11. Kardasisa, P., Loucopoulos, P.: Expressing and organising business rules. Informa. and Software Tech. 46, 701–718 (2004)
12. Ambler, S.W.: Agile Modeling. Wiley, New York (2002)
13. Bertino, E., Catania, B., Zarri, G.P.: Intelligent Database Systems: A Synopsis. Addison Wesley, New York (2001)
14. Sonar, R.M.: Integrating Intelligent Systems using an SQL-database. Expert Syst. Appl 17, 45–49 (1999)
15. Wu, S., Chang, C.-S., Ho, S.-H., Chao, H.-S.: Rule-based intelligent adaptation in mobile information systems. Expert Syst. Appl. 34, 1078–1092 (2008)
16. Eberhart, A.: Towards semantically enriched business logic. Electron. Commer. R. A. 2, 288–301 (2003)

Semantically Sensitive Execution of Relational Queries

Tomasz Wardziak and Maria E. Orłowska

Polish-Japanese Institute of Information Technology, Warsaw, Poland
wardziak@pjwstk.edu.pl, orlowska@pjwstk.edu.pl

Abstract. This paper presents concept of an alternative, semantically richer execution of queries to a relational database. We outline the formal extension of Relational Algebra to form Semantic RA that serves as the formal foundation for construction of new facility called Semantic SQL (SSQL). SSQL is a modification of SQL allowing matching and subsequently elimination of semantic duplicates. This new facility has no impact on targeted database content and existing applications. Queries expressed in SQL are executable as before and SSQL becomes only an optional alternative. A brief description of an implementation of SSQL for SQL Server 2005 concludes the paper.

Keywords: data quality, semantic equivalence, data mismatch, semantic SQL.

1 Introduction

The problem of data quality persists, and continues to be widely recognised [1]. Research works, as well as product developments, follow the urgent call to prevent semantic/syntactic data contamination, and provide reliable data quality control methods.

Historically, one of the first successes in data integrity control was the invention of referential integrity enforcement in relational database systems [17]. The generality of this solution, based on a formal definition of a class of constraints, made this data management concept uniformly applicable (independent of application domain), thus eliminating large numbers of data integrity errors. Now, it is regarded as a database industry standard feature to offer the integrity check mechanisms for most data processing products.

The published literature in this area can be classified into problems dealing with centralised systems (storing data in one logical database but operated in a multi user environment), and those where the data is distributed at multiple sites either generated locally or imported (inserted) from multiple independent sources to the integrated structure, such as a data warehouse or a BPM engine, or a scientific grid.

For centralised systems, identified data quality problems can be listed as follows: duplicate removals/identification, concurrent data access control, missing/incomplete data for attributes with mandatory occurrence constraints, inaccurate data entry (errors in spelling), integrity constraints violation, and infringement of other conceptual constraints such as frequency occurrence, subset constraints and subtype constraints.

For systems operating on integrated data, all problems with maintenance of data quality listed above are evidently present too, but an additional class of difficulties are

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 436–446, 2008.

also found. These include differences in decisions at the schema design level, different naming conventions adopted and conflicts of various local constraints. Difficulties in consolidation of data with semantic differences are well studied [6]. Initiatives on the development of solutions based ontologies (see e.g. W3C OWL & RDF) are worth mentioning in this regard.

Further, works on so-called 'approximate join', address some of the data discrepancy problems prevalent in data quality assessments. Results presented in [8], [11] show applicability and (lack of) a distance-based comparison between terms used as data values to selected applications. Approximate joins have attracted a lot of attention in the research community recently. It is a 'smart' invention based on the concept of a standard relational join but applied to the data of 'almost' match (not equal match as a natural join requires). In fact, it is only a version of a well known relational theory operation called theta-join. Our observations along the lines of practical applicability of approximate join versus merely being a computational challenge have been a genuine inspiration for this work.

The general problem of data quality has found a place in dedicated venues (see e.g. [3], [4]), and some methods have reached a mature stage of vendor investments. Many technologies (DataFlux, SAP MDM, Trillium Software) support various aspects of the data extraction, data transformation, enhancement and data loading. Terms such as CDI (Customer Data Integration), MDM (Master Data Management), EII (Enterprise Information Integration), IQ (Information Quality) have been added to the jargon of communication in the Enterprise Computing domain. However, they mainly deal with typographical issues and data enrichments by combining two or more data sources with a common (and we assume, correctly specified) reference point. For instance, for each item code the system automatically assigns an item name. Support is not available for the end-to-end process dealing with semantic data discrepancies designed specifically for the cleansing processes that ensure the desired level of data quality.

The main reason for this shortcoming is due to the fact that the data fusion process only begins with schema integration [5], and must be effectively followed by detailed data instances resemblance in order to reach data representation unification. Data mismatch (compared data is semantically equivalent but native representation of values of given attributes is different), by contrast to schemata mismatch, is frequently observed in practice, such as: different abbreviation conventions, different standards for data representation, different units and coding, etc., but under a common attribute name across many local data sources.

It is important to acknowledge at this point that the source of data quality problems is primarily found in human data entry error [2], or lack of knowledge manifested by lack of agreement in usage of terms to name attribute values. Managing the source of errors requires targeted organizational strategies and highly apt interfaces. However, it is evident that data entry errors can never be completely eliminated. Since the time distance between when the error is introduced to when it is detected may be significant, it is imperative that technological solutions "fix" the problem before it starts to impact on operational costs.

Ideal technology to remedy these problems must be able to access and extract data from additional sources, including legacy and non-relational data systems, and especially established databases for data quality control. It must provide an audit trail and reliable information on data lineage when testing the data to ensure its integrity, and identify all anomalies, incompleteness and inconsistencies. It must convert, cleanse, restructure, and consolidate the data as required. It must be capable of reconciling and validating the data in its final repository, providing analyses and reports at both a detailed and summary level. Finally, it must be computationally efficient, to support timely, cost-effective processing of data.

This paper introduces concept that can be used with multiple equivalences defined over each attribute providing a way to cooperate with many different databases and at the same time using original data in all databases involved without any data modification.

The query system for such language is based on Semantic Relation Algebra (SRA) that is equivalent in some form to the extended Tuple Calculus system. Due to the space limitation in this paper, the SRA is only briefly introduced followed by demonstrating how the queries expressed in SRA can be translated to this extended version of the SQL called Semantic SQL (SSQL).

To highlight the contribution and powerfulness of the introduced concept some characteristic features of the SRA are listed below:

- SRA is a generalisation of RA,
- All SRA's expressions are functions of traditional RA operations,
- SRA can be used selectively for an attribute or a set of attributes in a table,
- SRA forms the foundation to a new relational query language; SSQL that can be executed by an RDBMS without any system modifications,
- Semantic relationships between terms can be extended, shrink or modified without any other intervention than data entry to the database extension relations – no impact on the actual data,
- For one database, SRA may cooperate with multiple semantic extensions for as many attributes as required; one attribute extension per application,
- SRA can be used to determine hidden sub-typing relationships without any original data modification – this is a special case of semantic unification,
- Efficiency of SRA execution is built on the optimisation of RA deployed by RDBMSs,
- SRA offers enriched semantic integrity constraints enforcement - (uniqueness constraints checking for semantically equal terms, sub-typing and others).

Our goal is actual exploration and delivery of better solutions but surrounded with precise conditions that guarantee the data unification improvements if applied. There is a definite need to introduce a new data processing concept capable to unify semantically equivalent data but expressed differently (syntactically different) by independent systems (data entry facilities). In fact we need a complete set of new operations that could be called Semantic Relational Algebra (SRA) that forms a formal foundation to SSQL.

It is important to make it clear that without access to additional semantic matches (specially constructed to deal with data representation), unification of frequently used terms, and with different standards used to express the data values etc., there is no way to offer any guarantee as to the automatic data transformations being acceptable. It may sound rather pragmatic that we need to build 'schemes of transformations' but as we indicated above the subjectivity used by different users, applications and any data entry unities prevents any other reasonable and reliable solutions. Let us note that any judgement leading to the conclusion that the problem has been trivialised, is inaccurate here.

The greatest difficulty lies in positioning core solutions within practical expectations meaning that the current systems, without any modification to the data, will be able to process many different unification schemata as a plug-in service.

2 The Main Concept Introduction

For the completeness of this presentation and systematic introduction of the intended notation, we begin with a presentation of the traditional relational algebra that often serves as a query system for relational systems. Let $\mathbf{A = (A_1, A_2,..., A_p)}$ be a set of attributes called the universe. Let **D** be a set of domains, and let ***dom*** be a total function from A to D. Let $\mathbf{R = (R_1, R_2,..., R_p)}$ be a set of distinct relational schemas where R_i is a subset of A for $1\leq i \leq p$. Let $\mathbf{d = (r_1, r_2,..., r_p)}$ be a set of relations such that r_i is a relation on R_i, $1\leq i\leq p$. Let **Q** be a set of comparators over domains in D, including at least the equality and inequality comparisons for every domain.

The Relational Algebra (RA) over A, D, dom, R, d and Q is the following tuple **RA = { A, D, *dom*, R, d, Q, Op }** where **Op** is the set of operators union, intersection, difference, active complement, project, natural join, divide, renaming using attributes from A, select using comparisons from Q, logical connectives and theta-join using comparisons in Q.

An algebraic expression over R is any expression formed legally (according to the restrictions on operators) from the relations d and constant relations over schema in A, using the operators in Op.

2.1 Equivalence of Terms

To generalise the concept of RA we introduce a formal mechanism to deal with grouping of all semantically equivalent values of a given attribute A_i into distinct classes. Hence, for each attribute A_i from A, we include an equivalence relation E_i as follows: $E_i \subseteq$ [dom (A_i) x dom (A_i)] such that two values of attribute A_i are in relation E_i if and only if they are designated to be semantically equivalent.

Further, we name by E set of all E_i for all $1\leq i\leq p$, so E = (E_1, E_2,…, E_p). We introduce notation for set of equivalence classes generated by each E_i as EQC_i, EQC_i = (e_{i1}, e_{i2},…, e_{in}) where n depends from E_i. Clearly, two different attributes will most likely have a different cardinality of their EQC sets.

For the sake of simplicity of this presentation we focus now on only a single E_i for each Ai. However, another level of generalisation is possible, where many different equivalence relations can be defined for each single attribute.

Let us focus on one attribute A_i from A. The question arises how to express the imposed definition of semantically equivalent terms. Note that at any given time we may wish to modify the membership of a class by either removing some elements or adding new ones to reflect our application needs. Thus the maintenance of such representation of imposed classification is of consideration for purely practical aspects of our invention. Hence, for the efficiency of E_i maintenance, rather than storing all possible combinations of equivalent terms (a quadratic number of records and single modification always triggers many additional combinations to be maintained), we group all equivalent terms in equivalence classes and label them with a unique class name or number.

A natural way to represent such E_i for A_i relationship (in relational environment) is to build a relation v_i for each i, that represents association between the terms and its class identifier. For each i, v_i table has two attributes; ***term name*** called T_i with $dom(T_i) = dom(A_i)$, and ***class id*** called C_i, with $dom(C_i) = N$ where N is set of natural numbers. Hence semantic equivalence (or similarity) defined on attribute A_i is represented by relation v_i (T_i, C_i). Let further $V = (v_i$, for all $1 \leq i \leq p$) be the set of all v_is for a given $A = (A_1, A_2, \ldots, A_p)$.

A typical EQC_i will contain either a single value (when we don't recognise any other terms being equivalent with this value) or multiple terms when we wish to unify some terms. In the latter case it would be practical to select a single term that would be used in our query processing to identify any other term from its class. This is the way we introduce, as an optional reference, to a given e_{ij} from EQC_i a designated name – by selecting the preferred option of terms to be used to name all other equivalent terms from that class. To specify such representation we launch one more relation associated with (A_i, E_i); a table that keeps designated members' names from each class associated with this class identifier. We call this a master name for each class e_{ij}. Let w_i be such relation. Naturally, it has the same relational schema as v_i thus w_i (T_i, C_i). Clearly, the number of records in this relation is always equal to the cardinality of EQC_i defined as a partition of A_i.

For completeness, we name the set of all w_is for a given V by $W = ($ w_i, for all $1 \leq i \leq p$). Further on, relations v_i and w_i for A_i, we call a semantic extension of A_i. When applied to all attributes of a table $r(R) = r(A_1, A_2, \ldots, A_k)$ we define a semantic extension of r as follows $r ((A_1, v_1, w_1), (A_2, v_2, w_2), \ldots, (A_k, v_k, w_k))$.

One can make the following observation in regard to inclusions dependency between relation $\prod_{A1}(r)$ and its corresponding v_1; in general, the intersection of those relations should be quite sizable but none of them have to be included in another. This observation triggers a very important general issue. Basically, the question is how to effectively construct semantic extensions to attributes in our database. It would be unrealistic to expect that v_i will be constructed by a user wishing to explore some semantic operations on the data, especially not knowing the current content of the database. However, without user input to the definition of semantic equivalence of terms, it is also an unrealistic expectation to generate such dependencies fully automatically. We need to strike a right balance here, meaning, to only use the user's input where absolutely necessary.

A practical option, leading to smaller additional data sets (v_is tables), hence potentially faster in execution later, is to store only substantial equivalence – the classes with more than one element. This means that we only specify equivalence when required (by application) and the remaining values will be treated as individual elements as they are in the traditional RA executions. However, such a reduction of v_i size is overwhelmed by the additional complexity of our new relation expressions. Decisions about what option of database extension we choose have a big impact on some definitions of SRA expressions.

The v^*_i is a relation that includes all one-elementary classes as well as classes specified by the user. Let r $(A_1, A_2, \ldots, A_k)$ and A_1 classification defined by v_1 and w_1 be given. We consider here an extended version of r's schema by tuple's id (unique id maintained by the DBMS for each tuple in each relation), hence r schema is extended by one column as follows: r $(A_1, A_2, \ldots, A_k, id)$

$$v^*_1 (v_1) = v_1 \cup \{ (\prod_{A1}(r) - \prod_{T1}(v_1)) \otimes_{A1} \prod_{A1, id} (r) \}$$

Note that v^*_i depends only on the user specified v_i. From now on when we talk about database extension for attribute A_i we will think about v^*_i. We should also mention here about the relationship between relations from V and W. It should be clear by now that for each $1 \leq i \leq p$ v_i is a superset of w_i meaning that all records in w_i are also records in v_i for all i. It seems natural then to extend the notation of v^*_i to its subset relations w^*_i. To avoid unnecessary confusion we propagate this notation to the sets of those relations V* and W* respectively.

2.2 Generalisation of RA

Each operator in SRA has its corresponding counterpart in RA but expressed by an algebraic expression defined in RA on R, E, V and W such that when Ei is empty for all i then SRA = RA. Note here that if E is empty then V and W are empty too. It is worth mentioning here that 'empty' means only that there is no any unification of terms required, all attribute values theoretically are different so belong to different e_{ij}s for all attributes. It us our convention here to call such E an empty relation while in fact, it is a well defined equivalence relation with as many EQC elements as distinct values in $r.A_i$ for each $1 \leq i \leq p$. The reason for calling such E empty becomes apparent when we introduce more details about the SRA expressions construction.
Let us summarise the components of SRA:

$$SRA = (A, E, D, \mathit{dom}, R, V^*, W^*, d, Q, Op, SOp) = RA \cup (E, V^*, W^*, S0p)$$

where SOp are corresponding 'semantic aware' versions of Ops.

2.2.1 Semantic Projection Operator - SProjection

Intuitively, the projection of r onto X is the relation r'(X) is obtained by striking out columns corresponding to attributes in r-x and removing duplicate records in what remains.

Let X=A1. SProjection is an n-ary operator taking as its arguments; relation $r(A_1, A_2)$ and its extension v^*_1 and w^*_1. Since v^*_1 and w^*_1 are both functionally dependent from A_1 , we propose to adopt the following schema of the notation in this particular case; $\prod''_{A1}(r ; A_1)$ where $\prod''$ indicate that this is SProjection and $(r ; A_1)$ its arguments are relation r with one attribute A_1 being extended.

For the general case, we propose to adopt the following notation and conditions for well-formulated SProjection:

$$\prod''_X (r ; A_1, A_2, \ldots, A_k) \text{ where } X \subseteq R \text{ and } (A_1, A_2, \ldots, A_j) \subseteq X$$

Definition of semantically sensitive projection SProjection for $X=A_1$ is as follows:

$$\prod''_{A1}(r; A_1) = \prod_{w^*1.T1} [(\prod_{A1}(r) \otimes_{A1=T1} v^*_1) \otimes_{v^*1.C1=w^*1.C1} w^*_1]$$

Note that if we will not introduce v*s and w*s the our expression in this case would be:

$$\Pi''_{A1}(r; A_1) = \Pi_{w1.T1} [(\Pi_{A1}(r) \otimes_{A1=T1} v_1) \otimes_{v1.C1=w1.C1} w_1] \cup [\Pi_{A1}(r) - \Pi_{A1}(v_1)]$$

2.2.2 Semantic Select Operator - SSelect

Select is a unary operator on relations. When applied to a relation r, it yields another relation that is the subset of tuples of r with a certain value on a specified attribute as defined in the predicate being a variable part of select specification. Let r be a relation on schema R. A_i an attribute in R, and a an element of dom(A_i).

For simplicity of this presentation, as before, we assume that relation r is defined on schema R = (A_1, A_2) and only attribute A_1 has its semantic extension defined. Then semantic select SSelect is an 3-ary operator on relations r(R) and its extension v^*_1 and w^*_1 . It yields another relation that is the subset of records of r_i with values on the specified attributes not only equal but all equivalent to and named using the master data. We will use $\sigma''_{A1=a}(r; A_1)$ to mark this version of SSelect where σ" indicates that this is SSelect.

Let us restrict this expression to a relation r(A_1,A_2) with two attributes only (the general form for any R is a simple extension and with some effort it can be generated).

$$\sigma''_{A1=a}(r; A_1) =\{ \Pi_{A1,A2} ([\Pi_{A2, C1}(K \otimes_{A1=T1} v^*_1)] \otimes_{C1} w^*$$

$$\text{where } K = \Pi_{A1,A2} [(\Pi_{T1} (\sigma_{T'1=a} (v^*_1)) \otimes_{C1} v^*_1) \otimes_{A1=T"1} r]$$

Note that if we use only multi-elementary classes as data in v_1 and correspondingly in w_1 then our basic expression is more complex as below;

$$\sigma''_{A1=a}(r; A_1) =\{ \Pi_{A1,A2} ([\Pi_{A2, C1}(K \otimes_{A1=T1} v_1)] \otimes_{C1} w_1)\}$$
$$\cup \{\sigma_{A1=a} [(\Pi_{A1}(r)- \Pi_{T1}(v_1)) \otimes_{A1=T1} r] \}$$

This quite complex expression caters for both cases – if we store one-elementary classes in v relations or not. The union of sub expressions guarantees correctness of evaluation in this respect. Further on we will adopted extended attributes' extensions v^*_1 and $w^*_{1.}$ It is worth to note that SSelect guarantees correctness even if one uses the attribute value that is not currently used in the database but there is record(s) that should be returned because it is (are) equivalent to the value we have chosen for the query definition.

2.2.3 Semantic Join Operator - SJoin

The standard Join is a binary operator for combining two relations. In general, Join combines two relations on all their common attributes. Clearly, semantic equivalence relation introduced on a single attribute's domain can be now extended to deal with values from any two attributes sharing the same domain. In fact our previously introduced semantic equivalence was a relation defined on a domain such that any two elements a and b from the domain were equivalent (a ≅ b) if and only if they were declared to be synonyms or due to other reasons we would like to unify their values; for instance elements from one subtype of entity type structure.

We consider two relations: r(A_1, A_2) and s(B_1, B_2), further let dom(A_1)=dom(B_1). Let E be equivalence relation defined on dom(A_1). The semantic extension of A_1 let be v^*_1 (T_1, C_1) and w^*_1 (T_1, C_1). Naturally, the extension on A_1 applies to B_1 whether it is defined or not but B_1 can't be subject on different classification. For that reason we avoid any confusion by using only index 1 in reference to the classification on A_1 . The structure of this computation is simple; we first remove all semantic duplicates from both tables in respect to the classified attributed and then perform traditional join between the tables.

$$r \otimes''_{(A1=B1;\,A1)}\, s = \{\Pi_{w^*1.T1,\,A2,\,B2}\,[\,\Pi_{w^*1.T1,\,A2}\,(\,(r \otimes_{A1=T1} v^*_1) \otimes_{C1=\,w^*1.C1} w^*_1)\,]$$
$$\otimes_{w^*1.T1=\,w^*1.T1}\,[\,\Pi_{w1.T1,\,B2}(\,(s \otimes_{A1=T1} v^*_1) \otimes_{C1=\,w1.C1} w^*_1)\,]\,\}$$

For completeness of this presentation we look at this operation while storing abbreviated form of attribute extension. In case where attribute extension is expressed only by v_1 and w_1 the Semantic Join Operator, SJoin defined on r and s can be expressed as follows:

$$r \otimes''_{(A1=B1;\,A1)}\, s = \{\Pi_{w1.T1,\,A2,\,B2}\,[\,\Pi_{w1.T1,\,A2}\,(\,(r \otimes_{A1=T1} v_1) \otimes_{C1=\,w1.C1} w_1\,)\,]$$
$$\otimes_{w1.T1=\,w1.T1}\,[\,\Pi_{w1.T1,\,B2}\,(\,(s \otimes_{A1=T1} v_1) \otimes_{C1=\,w1.C1} w_1)\,]\,\}$$
$$\cup \{(\Pi_{A1}(r) - \Pi_{T1}(v_1)\,) \otimes_{A1} r\,) \otimes_{A1=B1} (\Pi_{B1}(s) - \Pi_{T1}(v_1)\,) \otimes_{B1} s\,)\,\}$$

3 Semantic SQL

Semantic SQL (SSQL) based on SRA is SQL counterpart based on RA. To ensure high learning curve of SSQL, we specified syntax of this language to be as close to SQL's grammar as possible. To write semantically aware version of SQL query using our language it is required to simply add # sign in front of the attribute's name. Operator having this extended attribute as its argument will be treated by the execution mechanism as the semantic version.

According to the SRA presented in Section 2, SSQL must have two additional tables used to store semantic extensions for an attribute. To maintain already introduced terms we call these tables **V (term (T), class id (C))** and **W (master term (T), class id (C))**. V is used to group values of attribute A into equivalence classes and W is storing mater names for these classes.

For the brevity of this presentation we introduce only three main SQL operators in their SSQL version: SELECT, WHERE and INNER JOIN. Other types of queries can be formed based on the similar principles.

3.1 Semantic SELECT Operator

```
SELECT #column1 FROM table1
```

This general form of SSQL SELECT operator is transformed into following SQL query:

```
SELECT DISTINCT T
FROM (
 (SELECT column1, C
  FROM table1
  INNER JOIN v ON table1.column1=v.T)
  AS t1 INNER JOIN w ON t1.C=w.C)
```

3.2 Semantic WHERE Operator

```
SELECT * FROM table1 WHERE #column1='val'
```

This general form of SSQL WHERE operator is transformed into following SQL query:

```
SELECT *
FROM table1
INNER JOIN
(
  SELECT w.T AS cname, t1.T FROM w
  INNER JOIN
  (
    SELECT T,C
    FROM v
    WHERE C=(SELECT C FROM v WHERE T='val')
  ) AS t1
  ON w.C=t1.C
) AS t2
ON column1=t2.T
```

3.3 Semantic INNER JOIN operator

```
SELECT column1, column2, ...
FROM table1
INNER JOIN table2 ON #(column3=column4)
```

This general form of SSQL INNER JOIN operator is transformed into following SQL query:

```
SELECT DISTINCT column1, column2 FROM
( ( SELECT column1,column3,C FROM table1
    INNER JOIN v ON table1.column3=v.T
  ) AS t1 INNER JOIN w ON t1.C=w.C
)
INNER JOIN
( SELECT DISTINCT T as Tinner, column2 FROM
  ( ( SELECT column2,column4,C FROM table2
      INNER JOIN v ON table2.column4=v.T
  ) AS t1 INNER JOIN w ON t1.C=w.C )
) as p2
ON T=Tinner
```

4 Overview of SSQL Execution Environment – An Implementation

In this section we briefly present main technological and architectural assumptions taken during the implementation process of our idea.

Figure 1 presents architecture of our solution. It is based on Microsoft .NET's Data Providers mechanism and SQL Server 2005. Developers can seamlessly integrate our provider with their existing applications. After extending SQL queries with their

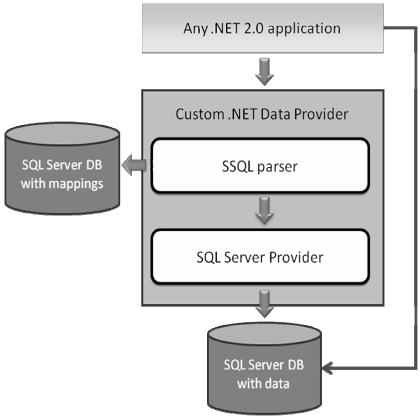


Fig. 1. Semantic SQL execution environment diagram

semantic versions applications can consume duplicate-free version of data the same way they worked with out of the box .NET data providers. SSQL data provider can be used also to process SQL queries, without having to change it during the execution. This enables developers to use SQL and SSQL the same way and treat semantic version of database query language as the extension to original SQL rather than its substitution.

SSQL parser is implemented using the visitor pattern. Every SSQL query passes by this mechanism and is evaluated against database containing tables with semantic mappings. Later it is transformed into corresponding SQL queries which are consumed by standard SQL Server Provider that is responsible for communication with source databases.

5 Conclusions

Presented concept is a powerful yet simple to use query extension to SQL language. SQL programmers will need little time to apply SSQL mechanisms in their current work. Existing .NET applications can be easily extended to process semantic queries thanks to the exchangeable Data Provider written for SSQL parser.

Ideas presented in this paper are foundations for multi-purpose query system. Flexibility of presented idea allows us to generalize presented concepts from relational system to XML data and also object oriented systems supporting multi-attributes equivalence classes.

Next step in our research and development efforts will be mechanism for integrating and unifying attributes from distributed database instances. Such environment will enable developers to seamlessly integrate data from many instances, not only unifying its schema, but also data semantics. It is crucial to ensure that integrated data is using the same meanings and our approach will allow to easily define semantic equivalences without changing single record in source databases.

References

1. Information Quality Research at MIT, http://mitiq.mit.edu/
2. Orr, K.: Data Quality and System Theory. Communications of ACM (1998)
3. CAiSE workshop on DQ, http://www.computing.dcu.ie/research/dataquality/diq/
4. ICDT workshop on DQ: http://www.dis.uniroma1.it/%7Edq/dqcis/
5. Halevy, A.Y., Ashish, N., Bitton, D., Carey, M., Draper, D., Pollock, J., Rosenthal, A., Sikka, V.: Enterprise information integration: successes, challenges and controversies. In: SIGMOD 2005: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, pp. 778–787. ACM Press, New York (2005)
6. Colomb, R.M., Orlowska, M.E.: Interoperability in information systems. Information Systems Journal 5(1), 37–50 (1994)
7. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. The VLDB Journal 10(4), 334–350 (2001)
8. Gravano, L., Ipeirotis, P.G., Jagadish, H.V., Koudas, N., Muthukrishnan, S., Srivastava, D.: Approximate string joins in a database (almost) for free. In: Proceedings of 27th International Conference on Very Large Data Bases, pp. 491–500. Morgan Kaufmann, San Francisco (2001)
9. Gravano, L., Ipeirotis, P.G., Koudas, N., Srivastava, D.: Text joins in an rdbms for web data integration. In: Proceedings of the Twelfth International World Wide Web Conference, pp. 90–101 (2003)
10. Gravano, L., Ipeirotis, P.G., Koudas, N., Srivastava, D.: Text joins for data cleansing and integration in an rdbms. In: Proceedings of the 19th International Conference on Data Engineering, IEEE Computer Society, Los Alamitos (2003)
11. Koudas, N., Marathe, A., Srivastava, D.: Flexible string matching against large databases in practice. In: Proceedings of the Thirtieth International Conference on Very Large Data Bases, Morgan Kaufmann, San Francisco (2004)
12. Ullmann, J.R.: A binary n-gram technique for automatic correction of substitution, deletion, insertion and reversal errors in words. The Computer Journal 20(2), 140–147 (1976)
13. Sadiq, S., Orlowska, M.E., Foulger, C.: Data Flow and Validation for Workflows Modelling. In: ADC 2004, IEEE Press, Los Alamitos (2004)
14. Bernstein, Y., Zobel, J.: Redundant documents and search effectiveness. In: CIKM 2005, pp. 736–743 (2005)
15. Godel, K.: http://mathworld.wolfram.com/GoedelsIncompletenessTheorem.html
16. Delphi Group. BPM 2003 Market Milestone Report (2003)
17. Date, C.J.: Referential Integrity. In: Proceedings of 7th International Conference on VLDB, Cannes, France (September 9-11, 1981)
18. Butler Group. Data Quality and Integrity – Ensuring Compliance and Best use for organizational data assets (February 2006)
19. Carr, N.: Does IT Matter? Information Technology and the Corrosion of Competitive Advantage. Harvard Business School Press (2004)
20. Data Unification with SAP Netweaver – Delivering a singly version of the truth, http://www.sap.com

From Business Rules to Application Rules in Rich Internet Applications

Kay-Uwe Schmidt[1] and Ljiljana Stojanovic[2]

[1] SAP AG, Research, Vincenz-Prießnitz-Straße 1, 76131 Karlsruhe
Kay-Uwe.Schmidt@sap.com
http://www.sap.com

[2] FZI Forschungszentrum Informatik, Haid-und-Neu-Straße 10-14, 76131 Karlsruhe
Stojanovic@fzi.de
http://www.fzi.de

Abstract. The increase of digital bandwidth and computing power of personal computers as well as the rise of the Web 2.0 came along with a new Web programming paradigm: Rich Internet Applications. On the other hand, powerful server-side business rules engines appeared over the last years and let enterprises describe their business policies declaratively as business rules. This paper addresses the problem of how to combine the business rules approach with the new programming paradigm of Rich Internet Applications. We present a novel approach that reuses business rules for deriving declarative presentation and visualization logic. In this paper we discuss complex event processing as an essential requirement for rule-enabled Rich Internet Applications, and introduce a rule-based architecture capable of executing rules directly on the client. We propose to use declarative rules as platform independent model describing the application and presentation logic. By means of AJAX we exemplarily show how to automatically generate client-side executable rules with the aid of Rich Internet Application design patterns.

Keywords: Rich Internet Application, Model Driven Architecture, Production Engine, Complex Event Processing, Event Condition Action Rules.

1 Introduction

Today's business world is characterized by globalization and rapidly changing markets. Thus in recent years business processes do not change yearly but monthly, the product lifecycle has shrunk from months to weeks in some industries and the process execution time has decreased from weeks to minutes as a result of the technological progress over the last few years. On the other side, the life cycle of IT applications stayed constant over time [1]. Business rules already proved their potential of bridging the gap between dynamic business processes and static IT applications. By declaratively describing the policies and practices of an enterprise the business rules approach offers the flexibility needed by modern enterprises. At the same time with the dawn of Web 2.0, a

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 447–458, 2008.

new technology for Internet applications appeared: AJAX [2]. Because of Web 2.0 and AJAX, Rich Internet Applications (RIAs) emerged from their shadow existence in the World Wide Web. AJAX, in contrast to Adobe Flash[1], now enables RIAs running in browsers without the need for any additional plug-ins. Several Web 2.0 applications use AJAX heavily in order to provide desktop-like behavior to the user. Now the time seems right for RIAs, because of the broad bandwidth of today's Internet connections, as well as the availability of powerful and cheap personal computers. Given those two trends observable in today's IT landscape, traditional ways of programming Internet applications no longer meet the demands of modern rule-enabled rich Internet applications. So, the strict distinction between declarative business logic and hard coded presentation logic holds no longer. As Web citizens are accustomed to highly responsive Web 2.0 applications like Gmail[2], Web applications based on business rules also have to provide the same responsiveness in order to stay competitive.

In this paper we propose a novel, rule-enabled architecture for RIAs. Our hybrid system architecture first automatically translates all relevant application rules into a client-readable format at design time. At runtime the application rules are executed directly on the client by a client-side rule engine. That enables a RIA to react directly to user interactions. The event patterns triggering the rules are found by a complex event processing unit. After identifying appropriate events, the application rules, in the form of event-condition-action-rules, are executed directly on the client. As a proof-of-concept and in order to evaluate the idea of using business and application rules as platform independent model, following the model-driven architecture paradigm, we prototypically realized a rule-enabled RIA using AJAX as client-side technology.

The rest of the paper is structured as follows. Section 2 gives an example motivating our work. In Section 3 we describe how to build application rules from business rules. Section 4 gives detailed insights into our rule-based architecture for RIAs. In Section 5 an overview of related work is given, and, finally, the paper closes with conclusions and prospects for future work.

2 Motivating Example

For motivating our work we chose an example from the financing sector. The example illustrating our approach is an online application for a loan. The use case is as follows: A person wants to apply for a loan from a bank. S/he visits the Web portal of that bank in order to fill in the online loan application. Figure 1 shows the form. The Web site offers four input possibilities: first, the name of the applicant; second, the amount of the requested loan; third the income of the applicant, and, finally, the kind of employment. The two buttons below the form submit or cancel the loan application.

The IT department of the bank decided to implement the online loan application as RIA in order to take advantage of the advanced visualization techniques.

[1] http://www.adobe.com/products/flex

[2] http://mail.google.com

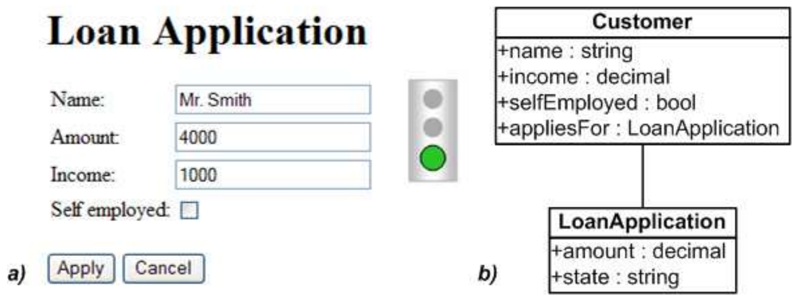


Fig. 1. Motivating example taken from the finance sector

The RIA shall give immediate feedback to the borrower signaling the probability of acceptance. Therefore, a traffic light was additionally introduced on the Web page. The lights indicate the status of the application for a loan. The *red* light signals a low or zero probability that the loan will granted. *Yellow* means that a clerk has to decide whether or not the loan application will be accepted. Finally the *green* light indicates that, based on the input data the loan will be granted in all probability. The traffic light shall change as the user fills in the online form without explicitly asking the server. That leads to a desktop-like behavior of the Web application.

The business logic of the application for a loan is well understood and written down as business rules, since they are subject to frequent changes. The RIA, and, especially, the manipulation of the traffic light, can reuse and build upon these business rules. The business rules shown in Example 1 declaratively represent the business logic behind a loan application[3]. The rules are written using the *IF/THEN* construct because of its simplicity and its commonness of use.

Example 1 (Business Rules).

1. IF C.income $\geq$ 1000 AND NOT C.selfEmployed THEN L.state = "accepted"
2. IF C.income $\geq$ 1000 AND C.selfEmployed THEN L.state = "to be checked"
3. IF C.income $<$ 1000 THEN L.state = "rejected"

Figure 1b) depicts the UML class diagram of the business objects (BOs). BOs are objects that encapsulate real world data and business behavior associated with the entities that they represent [3]. They are also called objects in a domain model. A domain model represents the set of domain objects and their relationships. The two BOs engaged in our example are *Customer* and *LoanApplication*. They are connected by the relation *appliesFor* which links one customer to one or many loan applications. The attributes of the customer class store customer specific attributes like name, income and employment status, whereas the attributes of the loan class hold loan specific data like the amount or the approval status of the application. A loan application can have the following statuses: *accepted*, *to be checked* or *rejected*.

[3] For the sake of simplicity we abstract from the amount of the loan.

The first rule states that if the borrower's income is grater then or equal to 1000 Euro and s/he is not self employed the loan will be granted in all probability. The second business rule states that if the income is greater then or equal to 1000 and s/he is self employed a clerk has to judge manually whether the loan will be granted or not. If the income is less then 1000 the loan will not be granted at all. In our example, all business rules are atomic. That means they are independent of each other and pairwise disjunct. That is of importance for our rule algorithms described in Chapter 5.

3 From Rule-Based Systems to Rule-Enabled Systems

Legacy rule-based Internet applications are based on the Web page paradigm[4] as depicted in the left graphic in Figure 2. User data are collected in forms on the client and are sent to the server by user request. On the server side a production engine processes the input data and derives actions for manipulating the application data. Based on the modified application data a new Web page is created. This dynamically created Web page is sent back to the client. Business rules in the back-end declaratively describe the business logic of the Internet application.

Rich Internet Applications (RIAs) break the Web page paradigm by introducing rich client-side functionality and asynchronous communication. The middlemost graphic in Figure 2 depicts the evolution of rule-based RIAs from common rule-based Internet applications. Up to date browsers provide a rich client engine capable of executing dynamic presentation logics. Together with the business logic, the production system stays on the server side but can be requested asynchronously. That is, business rules can fire without being explicitly triggered by a user request. RIAs have gained great momentum thorough AJAX, the enabling technology of Web 2.0 applications. Besides AJAX, other prominent members of the RIA enabling technologies are: Adobe Flex, Microsoft Silverlight[5], OpenLaszlo[6], to mention just a few.

Turning rule-based Internet applications into rule-enabled RIAs that benefit from the best of the two worlds is not trivial. Switching from the request/response communication of Internet applications relying on the Web page paradigm to the asynchronous communication of RIAs goes only half way. Although asynchronous communication with the Web server allows a RIA to reload only altered data rather then the page as a whole, as well as to pre-load chunks of data that might be good candidates for displaying next, the desired desktop-like responsive behavior is not achieved. This is because business rules and especially business rules concerned with the presentation layer are still evaluated on the server-side. Every user interaction, from pressing a button to hovering the mouse over an artifact on the Web site, must be processed on the server in order to let business rules encoding the presentation logic fire appropriate actions as reactions to the user's

[4] Every Web page in a series of pages is downloaded separately.
[5] http://www.microsoft.com/silverlight
[6] http://www.openlaszlo.org

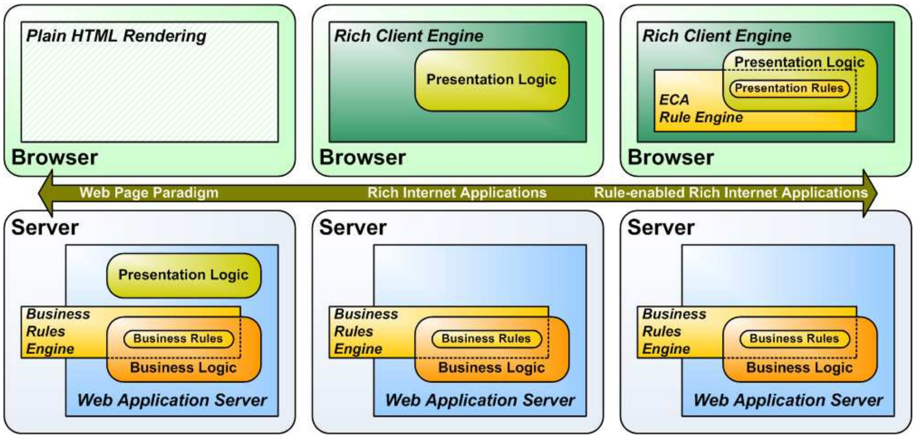


Fig. 2. Evolution of rule-based Internet applications

input. Also the advantage of the declarative character of rules is getting lost by only applying the rule paradigm to business logic and not to presentation logic. Presentation logic is also a good candidate for declarative modeling because it remains unchanged even for different platforms.

In order to have not only rule-based RIAs, but also rule-enabled RIAs, we postulate the following two requirements for application rules that enable declarative and model driven RIAs. Rationales like the need for a responsive user interface, as well as the need for a platform-independent model for the presentation logic, lead to the need for, firstly, declarative application rules that are, secondly, executable directly on the client within different execution environments (see the left graphic on Figure 2). Application rules describe in a platform independent manner the presentation logic formerly directly encoded in a programming language executable on a target platform like JavaScript for Web browsers. Application rules directly manipulate the user interface or change the state of the client-side application. They do not only change the surface of the application. Some user interactions may force the RIA to prefetch some application data, to send input data to the server or even to update internal stubs of business objects that are synchronized with their counterparts on the server-side after all user data are collected.

3.1 User Interface Design Patterns

Adopted from Alexander [4] a pattern describes a reusable solution to a problem that occurs multiple times. Patterns are prescriptive templates that help designers construct new instances. Patterns play an important role in model driven design as they allow the frequent reuse of problem solutions as well as they enable the declarative and, thus, platform independent description of them. User interface design patterns are frequently used for describing user interactions across the boundaries of different platforms. By not using concrete platform-specific input controls in application rules, the rules are able to serve as platform independent models, declaratively describing the application logic. User interface patterns are not only limited to input controls, but comprise also patterns like: architectural, structural and navigational patterns [5]. Besides user interface design patterns, there are some more design patterns crucial for RIAs. Those patterns describe the control flow of RIAs. The control flow logic is embedded in the application logic layer. According to [6] some control patterns are: RIA stub, fat client, predictive fetch etc.

We decided to model a RIA Design Patterns Ontology. We use ontologies mainly because of two reasons. Firstly, ontologies already have proven their applicability for managing large knowledge bases [7] and, secondly, rule languages like SWRL [8] fit nicely into the ontology paradigm. The RIA Design Pattern ontology describes all the patterns related to RIAs, that is, user interface and control patterns. Instances of these patterns can be used by applications rules in order to form the platform independent application model.

3.2 Deriving Application Rules from Business Rules

The starting point for every RIA is the business logic. The business logic declaratively encoded into business rules coarsely forms the frame of the user interface for RIAs. But business rules are usually high-level and are not related to any user interface issues. On the other hand, application rules presenting the presentation logic have to control, on a fine grained level, complex user interfaces. Therefore, the first step in creating application rule sets is the analysis of the business rules and their related business objects. Based on this analysis, the user interface and the presentation logic in the form of declarative application rules can be designed.

The application rule in Example 2 is directly derived from the first business rule (Example 1) with the difference that it does not manipulate business objects but a graphical widget signaling the acceptance status of the loan application. The application rule uses three design patterns from the RIA Design Patterns Ontology, namely, *NumbersOfAnyTypeOrFormat*, *BinarySelection* and *TrafficLight*. Although not necessary from a logical point of view, the type of the design patterns are listed explicitly in the rule for the sake of understandability. *NumbersOfAnyTypeOrFormat* and *BinarySelection* are subtypes of *InputControl* and *TrafficLight* is a subtype of *OutputWidget*. *NumbersOfAnyTypeOrFormat* is a design pattern for the input of numbers of any type or format. A concrete realization could be a common text field or a spin box. The design pattern *BinarySelection* allows the input of one of two options. For the latter, concrete realizations could be among others: a check box, two radio buttons or a toggle button. The output design pattern *TrafficLight* shows the status of the loan application and could be realized as traffic light or label.

Example 2 (SWRL Application Rule).

NumbersOfAnyTypeOrFormat(Income) & hasValue(Income, ?x) &
swrlb:greaterThanOrEqual(?x, 1000) & BinarySelection(SelfEmployed) &
hasValue(SelfEmployed, false) → TrafficLight(Status) &
green(Status, true) & yellow(Status, false) & red(Status, false)

3.3 The Need for Complex Event Processing

Graphical user interfaces are event driven. Every user interaction triggers an event that can be caught and processed. That event mechanism immediately suggests itself as a facility to evaluate only rules that are directly affected by user interactions. Application rules can be linked to triggering events. That strong link-up of rules, user interface patterns and events leads to a tremendous reduction of the evaluation time, as only those rules are evaluated that directly handle raised events. As the concrete form of events differs across platforms, we organized them hierarchically in the RIA ontology as design patterns as well. This allows their platform independent use in SWRL rules. So, for instance, in our motivating example (see Example 1), all application rules could be associated with a Change event pattern that gets raised whenever the value of one of the involved input controls changes. As the first two Business rules evaluate two input controls they also must be linked with two Change events that are related by a logical OR. That means the rules are evaluated whenever the value of one of the two input controls changes. The third rule only needs to be associated with one event, as it only evaluates the input of one input control. Complex events are composed of atomic or again complex events. Atomic events are events natively supported by the target platform. So as not to be dependent upon the platform, atomic events are also modeled as design patterns. The detection of such complex events is the task of a complex event processor.

Figure 3 shows a combined approach for deriving application rules from business rules, business objects as well as an ontology describing RIA design patterns like user interface design patterns, event design patterns and an event detection language. For a rapid design of application rules we strongly recommend a graphical design tool. This design tool should allow users to derive application rules from business rules, to design application rules from scratch, to define complex event patterns from simple event patterns and to link event patterns to rules.

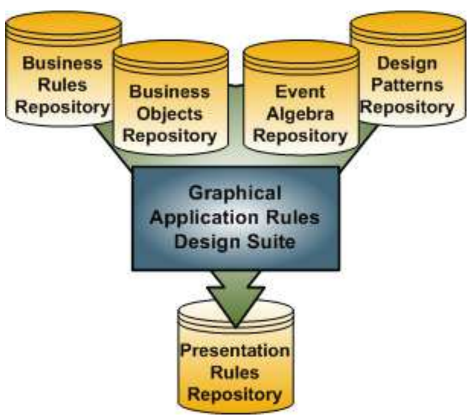


Fig. 3. Design of application rules

4 Architecture

We suggest a two level architecture for rule-enabled RIAs. The first level is the design-time level and enables the translation of the platform independent model into a platform specific model making use of the platform definition model. The second level executes the translated application rules at run-time on the client. When the business rules or the application logic alter, only the declaratively described rules need an adjustment. Then the rules must be translated again into platform specific models. Thus the rules become available to the RIA without any additional coding.

4.1 Design-Time Architecture

Following the model driven architecture paradigm, separating the design from the destination platform, the design-time architecture is responsible for generating executable code from the declarative application rules as well as from the associated design patterns. The logical design-time architecture is depicted in Figure 4. The platform independent model is declaratively given by the application rules. The application rules refer to user interface, event and other RIA design patterns stored in the RIA Design Patterns Ontology.

The *Translator* component in the middle of the picture carries out the entire work at design-time. It reads the application rules and a platform definition model, for instance, that of Silverlight. Afterwards it translates the whole platform independent model into the platform specific model of Silverlight with the aid of the platform specific model. In the platform specific model instructions are stored about how to replace, for instance, the *NumbersOfAnyTypeOrFormat* design pattern with an platform specific control. In the Silverlight case this might be the *TextBox* control from the *Windows Presentation Foundation Everywhere*, the graphical subsystem feature of Silverlight. Additionally event handlers, data structures for complex event processing as well as data structures suitable to the implemented evaluation algorithm of the rule engine are constructed.

All the constructed and translated platform specific artifacts are stored in the Compiled Rules Repository. This repository holds the executable application logic code, which is distinct for the different target platforms. During run-time

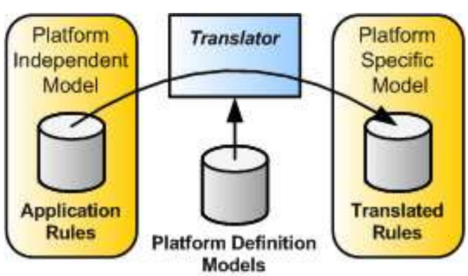


Fig. 4. Design-time architecture

the executable code can be asynchronously loaded by the client, and can be executed by the client-side rule engine.

4.2 Run-Time Architecture

The second level of our logical system architecture is the run-time architecture. Figure 5 depicts the client-side components of the run-time architecture. The server-side components are skipped for the sake of clarity. The software components of the run-time architecture carry out the application logic generated from the declarative application rules during design-time.

The application logic is transferred to the client together with the content data in response to the first initial user request. In the first preprocessing step the CEP Unit responsible for detecting complex events is initialized and, in a second step, the appropriate event handlers are set. As the events are linked to rules and controls in the application rules, this information needs to be extracted from them. Alternatively, the complex events could already be distilled from the rules beforehand during design-time. As complex events are not issued directly by user interface widgets the CEP Unit has to register for each atomic event contained in complex events.

Then the user interacts with the portal. He/she fills in forms, navigates through the site, goes back, searches for terms and so on. All those interactions trigger events like mouse movements in the appropriate controls. The CEP Unit handles all atomic events to which it has subscribed in advance (step three). Based on the directives of the event detection algebra, it tries to identify complex patterns from the event stream. After detecting a complex event, the associated rules are evaluated by the client-side rule engine. This is step four in Figure 5. In step five the condition parts of the rules are evaluated, if there are any, using common evaluation strategies of rule engines like Rete [9] or sequential algorithm [10]. If more than one rule is matched during the evaluation phase, they

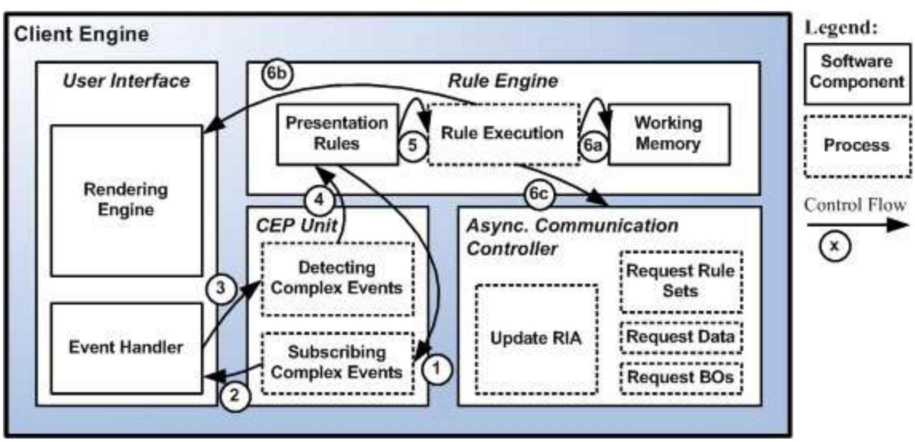


Fig. 5. Run-time architecture

are put into a conflicting set also called the agenda. Based on a conflict resolution strategy, one rule is picked and executed.

The execution of a rule can have manifold actions which are marked as 6a to c. In step 6a a rule manipulates the status of the application. The status of the application is maintained in working memory. In a nutshell, the working memory consists of an arbitrary amount of local object variables. Further a change to the working memory can trigger additional rules that are not explicitly bound to any complex event pattern. These rules are conventional production or condition action rules (CA rules). A rule can also manipulate the user interface directly as depicted in step 6b. By this means, application rules can respond to user interactions immediately without an explicit server request. These rules are the guarantors of a responsive user interface. Any user interface manipulation can issue additional atomic events that might be recognized by the CEP Unit as parts of complex events. New rules can be triggered. So the rule execution in step 6b can trigger additional rules over the event detection mechanism. The last possible action of a rule execution is depicted in step 6c: The invocation of the Asynchronous Communication Controller (ACC). The ACC is responsible for loading new rule sets, for prefetching content data as well as for synchronizing with the BO's on the server-side. As a direct byproduct of prefetching data and synchronizing with the server, the ACC can alter the user interface.

5 Related Work

Rule-enabled Rich Internet Applications seems to be a new and novel approach, as we could not find related work on this topic. Nevertheless, there exists already a reasonable amount of work addressing subtopics of our approach. Carughi and colleagues [11] describe RIAs as reactive systems where the user interface produces events. They use complex event processing in conjunction with server push technologies, but not for triggering application logic formulated in declarative application rules, that can be executed directly on the client. In their work complex events trigger some kind of server-side logic. They also do not address how complex events can be detected on the client-side.

The principles of complex event processing for reactive databases are well understood since the mid-1990s. Chakravarthy et al. [12] outline an expressive event specification language for reactive database systems. They also provide algorithms for the detection of composite events and an architecture for an event detector along with its implementation. Our work in the field of complex event processing relies greatly on their work and the work done in [13,14,15]. We propose to use the work done in these papers, lift them up to the client and enhance SWRL with concepts for detecting complex events. Recently some effort was undertaken to broaden RuleML[7] to a event specification language. As a result Reaction RuleML[8] [16] incorporates nicely different kinds of production,

[7] http://www.ruleml.org/

[8] http://ibis.in.tum.de/research/ReactionRuleML/

action, reaction, complex event processing and event logic rules into the native RuleML syntax but fails to support OWL ontologies.

Seffah and Gaffar [17] propose the usage of design patterns for model-based user interface engineering. They use design patterns in order to overcome limitations in model-based approaches for user interface engineering as the lack of reuse of best design practices. Our approach goes one step beyond as we additionally use declarative rules for describing the behavior of the design patterns at run-time. In [18] an design patterns ontology is introduced using Semantic Web concepts that turns informal patterns into formal representation capable of supporting systematic design methods. We designed our RIA Design Patterns Ontology based on the work done in this paper and refined the ontology to meet our needs.

6 Conclusions and Future Work

In this paper we presented a novel approach of using declarative application rules as a model for RIAs. The declarative application logic can be easily changed by rewriting the rules. Then the rules can be translated without additional coding into a format executable by arbitrary target systems like AJAX, Silverlight or Flex. We outlined the need for a declarative description of the application logic for RIAs and showed how SWRL rules in conjunction with design patterns can serve as a platform-independent model. Based on this model we were able to generate platform-specific rules for AJAX that can be evaluated at run-time by a client-side rule engine. So far we managed to test our ideas based on a simple example and the sequential algorithm. Several research challenges and future works remain open. Currently we are enhancing our solution by a *CEP Unit* in order to deal with complex events and we plan to implement a client-side version of Rete in JavaScript and Silverlight. We will enhance the design-time rule translator and will implement a generator for Silverlight. Finally, we will investigate the possibilities of enriching SWRL with an event detection algebra.

References

1. MacDonald, N.: Strategies for business growth. In: Gartner Symposium ITXPO (2002)
2. Jesse James Garrett. Ajax: A new approach to web applications (2005), http://www.adaptivepath.com/publications/essays/archives/000385.php
3. Heidasch, R.: Get ready for the next generation of sap business applications based on the enterprise service-oriented architecture (enterprise soa). SAP Professional Journal 9, 103–128 (2007)
4. Alexander, C., Ishikawa, S., Silverstein, M., Jacobson, M., Fiksdahl-King, I., Angel, S.: A pattern language. Oxford University Press, Oxford (1977)
5. Tidwell, J.: Designing interfaces, 1st edn. O'Reilly, Sebastopol (2006)
6. Mahemoff, M.: Ajax design patterns, 1st edn. O'Reilly, Sebastopol (2006)
7. Fensel, D.: Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. Springer, New York (2003)

8. Horrocks, I., Patel-Schneider, P.F., Boley, H., Tabet, S., Grosof, B., Dean, M.: Swrl: A semantic web rule language combining owl and ruleml. Technical report, W3C Member submission 21 may 2004 (2004)
9. Forgy, C.L.: Rete: a fast algorithm for the many pattern/many object pattern match problem. Artificial Intelligence 19, 17–37 (1982)
10. Berstel, B., Bonnard, P., Bry, F., Eckert, M., Patranjan, P.-L.: Reactive rules on the web. In: Antoniou, G., Aßmann, U., Baroglio, C., Decker, S., Henze, N., Patranjan, P.-L., Tolksdorf, R. (eds.) Reasoning Web. LNCS, vol. 4636, pp. 183–239. Springer, Heidelberg (2007)
11. Carughi, G.T., Comai, S., Bozzon, A., Fraternali, P.: Modeling distributed events in data-intensive rich internet applications. In: Benatallah, B., Casati, F., Georgakopoulos, D., Bartolini, C., Sadiq, W., Godart, C. (eds.) WISE 2007. LNCS, vol. 4831, pp. 593–602. Springer, Heidelberg (2007)
12. Chakravarthy, S., Krishnaprasad, V., Anwar, E., Kim, S.K.: Composite events for active databases: Semantics, contexts and detection. In: Bocca, J.B., Jarke, M., Zaniolo, C. (eds.) 20th International Conference on Very Large Data Bases, September 12–15, 1994. Santiago, Chile proceedings, pp. 606–617. Morgan Kaufmann, San Francisco (1994)
13. Chakravarthy, S., Mishra, D.: Snoop: An expressive event specification language for active databases. Data Knowl. Eng. 14(1), 1–26 (1994)
14. Papamarkos, G., Poulovassilis, A., Wood, P.T.: Event-condition-action rule languages for the semantic web. In: SWDB, pp. 309–327 (2003)
15. Alferes, J.J., Tagni, G.E.: Implementation of a complex event engine for the web. In: SCW, pp. 65–72. IEEE Computer Society Press, Los Alamitos (2006)
16. Paschke, A., Kozlenkov, A., Boley, H.: A homogenous reaction rules language for complex event processing. In: International Workshop on Event Drive Architecture for Complex Event Process (2007)
17. Seffah, A., Gaffar, A.: Model-based user interface engineering with design patterns. Journal of Systems and Software 80(8), 1408–1422 (2007)
18. Henninger, S., Ashokkumar, P.: An ontology-based infrastructure for creating software pattern languages. In: 2nd International Conference on Design Science Research in Information Systems and Technology, Pasadena (May 2007)

Enabling Context-Based Cooperation: A Generic Context Model and Management System

Christian P. Kunze, Sonja Zaplata, Mirwais Turjalei, and Winfried Lamersdorf

Distributed Systems and Information Systems
Computer Science Department, University of Hamburg
Vogt-Kölln-Str. 30, 22527 Hamburg, Germany
{kunze,zaplata,9turjalei,lamersdorf}@informatik.uni-hamburg.de

Abstract. In order to realise complex service-based applications on system platforms for context-aware ubiquitous computing environments, *mobile processes* have been introduced to support cooperation among (mobile) devices by exchanging and executing arbitrary (business) processes. In such a view, middleware platforms that support the execution and migration of mostly a priori unknown processes need a generic and also application-independent context system. Accordingly, this paper presents an approach for a generic context model and management platform to support such *context-based cooperation* as currently developed and used in the project DEMAC (*Distributed Environment for Mobility-Aware Computing*).

1 Introduction

Mainly due to the technical progress in processor and network technology, ubiquitous computing environments are becoming more and more reality. One of the main characteristic of such environments is the mobility of users, devices, and even application code. Under such conditions, mobile applications and supporting platforms frequently have to bridge the gap between required and provided resources and capabilities at any specific place and/or time.

In order to narrow this gap, devices of respective mobile vicinities can *cooperate* by sharing their resources for executing mobile applications. Today however, in most cases, mobile applications are still restricted to the capabilities of those device(s) they were initiated on. Other resources which are, e.g., potentially available from other devices remain still inaccessible for any dynamic adjustment of the mobile application. In consequence, this also limits the complexity of applications and tasks to the initialing device's capabilities. But, in order to fully realise the vision of ubiquitous computing [16], even much more complex and also unknown tasks (and thus more generality) have to be supported by advanced mobile and context-aware systems. *Mobile processes* represent such complex application tasks and use context knowledge to execute and migrate (business) processes in order to increase the likelihood to finish the task successfully by integrating the capabilities of different mobile nodes of the vicinity

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 459–470, 2008.

(cp. [8] for a detailed introduction). As use of these processes leads to real cooperation among the participating mobile systems, applications implemented as mobile processes are able to realise a form of *context-based cooperation* [9]. However, a supporting context-aware middleware for this new class of applications has increased requirements for the underpinning context model and management system as compared to more traditional ones. Accordingly, this paper identifies such requirements and presents an approach to model and use context for a mobile system supporting the execution and migration of mobile processes.

The following subsections of the paper first motivate the need for a generic context model and management system for mobile processes. Afterwards, basic requirements of an underpinning context-aware middleware platform to support such processes are identified. Section 2 then addresses related work and section 3 presents the coarse architecture, the proposed context model and the corresponding context management component. Section 4 refers to the evaluation within the DEMAC middleware before section 5 concludes the paper with a summary.

1.1 Executing Service Compositions with Mobile Processes

A *mobile process* is a goal-oriented composition of arbitrary services and manual tasks which may span several heterogeneous mobile and static devices, users, and services. In order to (potentially) use all the capabilities and resources provided in its entire local and remote environment, such a process can *migrate* to other devices, e.g. to share the functionality provided by these nodes. In order to ensure the user's goals even on foreign devices, non-functional requirements can be specified to restrict participating services, users and devices. The required activities within such processes are determined by abstract service classes to refer and identify applications and services in a technology-independent way. This is done because the type of services and applications provided in the environment at runtime can not be determined at designtime - but result from the available resources of the context during execution. The same argument holds for the description and the management of context information: Because a mobile process can involve arbitrary tasks and heterogeneous devices - and therefore also different

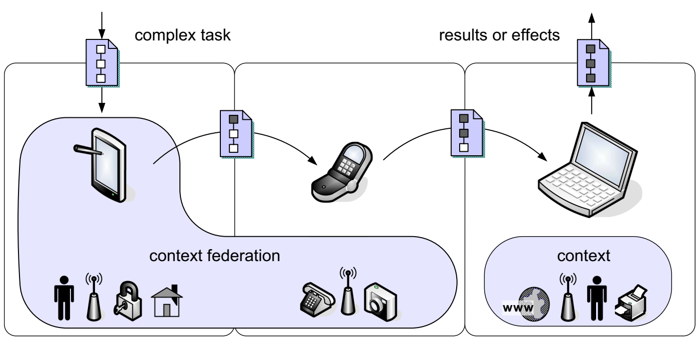


Fig. 1. Context-based Cooperation for Mobile Processes

contexts which cannot be determined in advance by the executing device - the context model and context management system of a supporting middleware infrastructure has to be generic and application-independent even at runtime. Figure 1 shows an (abstract) mobile process migrating three devices in dependence of the discovered context. As long as the process engine of a device is able to bind local or remote services to its current activity, it is responsible for the mobile process. However, in cases of failures or lack of respective service instances the engine has to try to find other devices to execute the mobile process and to transfer the remaining process and its execution to one of them. As the initiator of the process (in this case the user of the PDA) is interested only in the effects of the process execution, there is no need to return the results to this participant. A concrete example of such a mobile process can be found in [9].

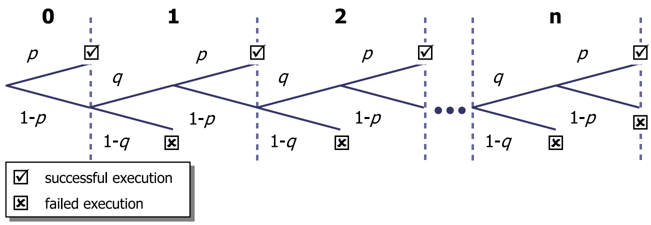
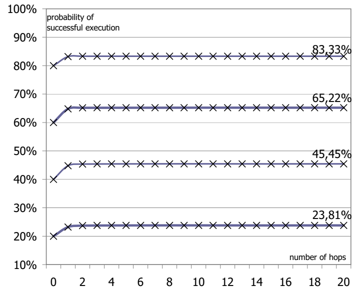


Fig. 2. Probability tree of successful execution

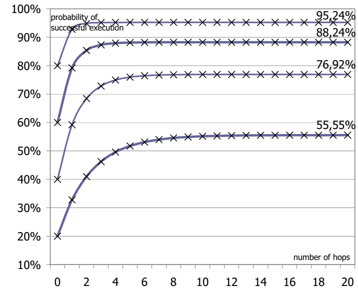
The procedure of migration opens up a new vicinity to search for other and maybe more suitable devices and is determined by the heterogeneity of the vicinity. Thus, the likelihood to finish the overall task successfully is increased in the following way (as depicted in figure 2): The upcoming task can either be directly executed by the current device itself or, otherwise, has to migrate to another device. Accordingly, let p denote the probability of a single device being capable of executing the current task. In extension, let q denote the probability of migration. Without restricting generality, let furthermore p and q be equal for all devices of the mobile vicinity, and n be the number of hops caused by migration. Equation 1 summarizes these observations by calculating the probability as a converging geometric series as the likelihood of a successful execution of the task *anywhere* in the mobile vicinity. Some exemplary values are presented in figure 3 showing the probabilities of successful execution with a migration probability of 20% and 80%. As to see, the estimated probability of a successful execution increases considerably already after only a few number of hops.

$$P_{Success} = p \sum_{i=0}^{n} ((1-p)q)^i = p \frac{1 - ((1-p)q)^{n+1}}{1 - ((1-p)q)} \tag{1}$$

In summary, the probability of successful execution – involving arbitrary tasks and heterogeneous devices – is influenced by the migration which is itself dependent on the potential to detect and integrate heterogeneous context dynamically. However, since such different context cannot be determined in advance,



(a) Migration (q=20%)



(b) Migration (q=80%)

Fig. 3. Estimated probability of successful execution

the context model and context management system of a supporting middleware infrastructure has to be *generic* and *application-independent* even at runtime – which is hardly supported by traditional context-aware middleware systems. The following section therefore presents basic requirements for such a generic context-aware middleware approach for mobile processes.

1.2 Requirements Analysis

In order to facilitate the exchange of context information and interaction between heterogeneous mobile devices, relevant *context information* has to be mapped to an adequate data structure and thus be provided in a standardized way (1). Therefore, on the one hand, context information has to be represented by an adequate *uniform and extensible model* (2) which allows standardized interaction between mobile devices. On the other hand, techniques for context management, such as the discovery, evaluation and administration of relevant data, are required. To ensure reusability and extensibility, the context model and the context management system should be *conceptually decoupled* (3). Additionally, the entire context system has to be independent of platforms, communication protocols and programming languages to allow a *flexible integration* into existing mobile middleware (4).

A main aspect of cooperation is the *exchange of distributed context information* (5). Thus, not only the local context but also the contexts of other mobile participants in the vicinity are relevant. Therefore, the middleware should define a protocol to exchange context information and to provide efficient mechanisms to access the context of cooperating devices. Because different devices can include different contextual data, the context model has to be able to represent different categories of context information (e.g. environment, activity, identity or time). Therefore, it is necessary to *abstract* from the structure of contexts and their actual representation, e.g. the local time, temperature or velocity (6). Consequently, the management system has to support the *transformation* of context information to equivalent data formats on the basis of semantic knowledge (7).

In order to enable proactive and ad-hoc reactions to relevant changes of context, the middleware needs to observe the context model and *notify registered applications in case of specified changes* (8) in the overall context or of single attributes. Other important requirements are derived from considering constraints resulting from mobility. The scarcity of resources of mobile devices requires a *lightweight and robust system architecture* (9). This means, in particular, that the composition of middleware services has to be *flexible and scalable* (10) in order to allow, for example, deactivation of temporarily unused modules. Likewise, the selection of relevant and situation-specific context information requires custom *filter mechanisms* (11). Concerning the context model, the semantic separation of *high-level, low-level, dynamic and static context information* facilitates the runtime management (12): While dynamic context data is subject to frequent changes and has to be updated periodically, static context does not need costly refreshing procedures at runtime. Furthermore, non-functional aspects such as *security and privacy concerns* (13) as well as the specification of quality parameters (14) for the potentially imprecise data play an important role (*Quality of Context* [2]).

2 Related Work

As shown, context-based cooperation of services requires a generic context model as well as a flexible management system to support arbitrary applications defined as a mobile process. Respectively, this section presents an overview of related context modelling approaches and existing frameworks to support context management on mobile systems. A rather simple and thus lightweight approach of structuring context information is realized by the *key value model* [13]. To provide context information, the value of the relevant context parameter - like the current position - can be stored e.g. to the key of one respective environment variable [12]. As a consequence of its simple representation, this model is not adequate to structure composed or high-level context information. In contrast to that, *markup schema models* represent context data in more complex hierarchical textual structures [13]. Examples are the *Comprehensive Structured Context Profile (CSCP)* [6] and the *CC/PP Context Extension* [7]. Related to this markup approach, an *object oriented model* encapsulates context information also hierarchical but as classes. A more formal description of context can be derived by *contextual reasoning or inferencing* in logic-based approaches [10] or can be realized by *ontological models*, which introduce methodologies for a normalized domain-specific knowledge representation. Examples are *Aspect Scale Context Information (ASC)* [10] and *CONtext ONtology (CONON)* model [15].

These models are used, e.g. in the *Context Toolkit* [11] which is a management framework based on an object-oriented architecture. It uses *context widgets* to abstract from sensor-based raw data, provides reusable components to access context information, and supports the exchange, persistence and logging of context data. Although the widget approach is a suitable solution to abstract from low-level in-formation and to provide reusable components, there are no concepts about how to exchange context information in a distributed mobile environment. In contrast, the *Hydrogen* framework [14] was exclusively developed for mobile applications and is therefore characterized by a very lightweight architecture. However, communication between applications, middleware and sensor network is restricted to TCP/IP-based protocols. Furthermore, a transformation mechanism to obtain different but equivalent representations of context information is not available. The *Java Context Awareness Framework (JCAF)* [1] is a compact event-based system to support the development of context-aware platforms. It consists of an extensible API to model application-specific context information, and of a distributed infrastructure system, which provides cooperating context services in order to manage the context model. However, a clear separation between context model and management system is not consequently pursued. Furthermore, the communication between services and clients as well as among the clients themselves is based on Java RMI and, in addition, context information is modelled as serializable Java objects and thus JCAF cannot be considered to be totally platform-independent.

Although a detailed discussion of the whole range of related work is out of the scope of this paper, in general, it can be derived that *peer-to-peer* and *ad-hoc* infrastructures of portable devices form the most challenging environment for mobile applications and, consequently, systems which move at least important parts of the context computation into supporting infrastructure - such as *Nexus* [4] or *Solar* [5] - and relatively heavyweight ontological approaches - such as *SOCAM* [15] or *CoBrAs* [3] - are not considered in this paper. Table 1 evaluates (due to space restrictions only some) selected contemporary context-aware middleware approaches and the proposed system according to requirements for supporting context-based collaboration and outlines also exemplary deficiencies. In

Table 1. Analysis of Context-Aware Middleware Platforms

	Context Toolkit	Hydrogen	JCAF	DEMAC Approach
(1) Provision of Context Information	+	+ +	+	++
(2) Extensibility and Reusability	+ +	+	−	++
(3) Separate Model and Management	− −	−	+	++
(4) Integration	+	+	−	+
(5) Distribution	−	+	+	++
(6) Mapping and Abstraction Level	+	−	+	+
(7) Transformation	+	− −	+	++
(8) Extended Notification Mechanism	−	−	−	+
(9)Robustness and Compactness	+	+ +	+	++
(10) Flexibility and Scalability	− −	−	−	++
(11) Filter Mechanisms	−	−	−	++
(12) High Level and Static Context	+	−	− −	++
(13) Security and Privacy	− −	−	+	−
(14) Quality of Context Parameter	−	−	+ +	++

++ supported; + partly supported; − marginally supported; − − not supported particular, the evaluated approaches have at least some deficiencies w.r.t. a clear separation between context model and management system which influences their flexibility and scalability. Furthermore, mechanisms to filter relevant context information and to realize extended notification to consider changes in the overall context remain unsupported. Therefore, all approaches presented there do not suffice to support highly dynamic context-based collaborations which can also determine relevant context parameters at runtime.

3 A Context Component for Distributed Mobile Systems

In line with the well-established tradition of "middleware-based" approaches to open distributed application development, system support for mobile, context-aware applications shall also be realised by respective *infrastructure components*. In such a view, figure 4 provides an overview of the coarse architecture of a *context (middleware) component*. As shown, the *raw data layer* represents a collection of heterogeneous services and components which provide low-level context information without any structuring. These are either interfaces to physical sensors of the mobile device itself or logical sensors based on (remote) services. As the components of the raw data layer are composed arbitrarily and can also change during runtime, context information can be achieved from different resources. Therefore, the *service management* of the *context layer* has to abstract from these single resources in order to overcome the heterogeneity of its underlying services. The context management component uses homogeneous wrapper services from the service management to administer the context model, e.g. to update context data on the basis of new raw data. Both components are encapsulated by the *context service*, which acts as a proxy to access the context layer.

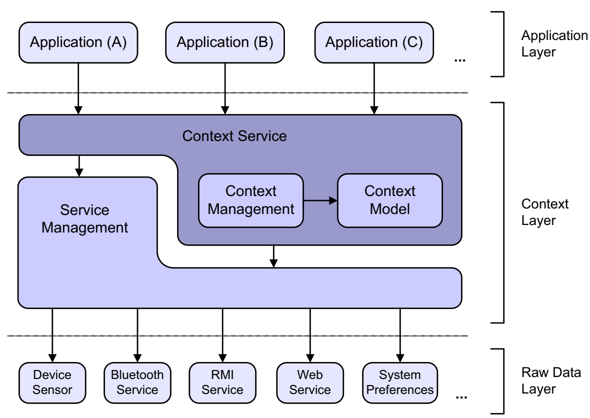


Fig. 4. A Context Component to Support Context Aware Middleware Systems

3.1 Generic Context Model

Although resources and services providing context are often heterogeneous and differ in representation and quality, the middleware should be able to allow for an intuitive and non-complex selection and composition of context information. Therefore, the context model has to provide standardized interfaces and formats to integrate different context resources. To illustrate the components of the presented context model and the inter-relationships, figure 5 shows a schematic diagram of this model: Here, an entity represents a single object which can aggregate several components to describe contextual information on a higher level as, e.g. a person with her location and activity or a meeting combining a set of persons. As can be seen in this example, complex entities can also aggregate other entities to describe their context recursively.

A single piece of context information within an entity is modelled as an *attribute*. Because the context information can be obtained by heterogeneous services with different characteristics, there are also different types of attributes. A *sensed attribute* indicates a dynamic context value, meaning that the context value changes relatively frequently which is typical for information from logical or physical sensors (e.g. the amount of free memory space of a mobile device). In contrast, a *defined attribute* models more static context information, which is unlikely to change during runtime, e.g. the identity of the user or the type and model of the mobile device. Finally, high-level context information, which is obtained by a combination or processing of several low-level context attributes, is described as a *deduced attribute*. The actual value of an attribute is modelled as the *data value* component and can include arbitrary simple or complex data types. Furthermore, each attribute provides a link to external semantic resources - e.g. an RDF file - as a unique identifier allowing to share context data between manifold applications more easily. Because contextual information can be imprecise, inconsistent, or incomplete, information about the quality of context helps to interpret the reliability of the data. Therefore, the component *quality constraints* contains a set of *quality parameters* as, for example, the refresh period of data, its precision, or its standard deviation.

On a higher level, the component *domain context* represents a self-contained and specifically demarcated environment which holds a limited number of relevant entities. This classification has the advantage to serve as a filter mechanism in a way that in a given context only that information is being selected, which is relevant to the respective application. As a result, the number of context properties discovered and held by the mobile device is as small as possible and the amount of necessary information is customized to the particular environment of the application.

The whole set of relevant domain contexts and information about the current domain is part of the *device context* which aggregates all context data of a single mobile system. The overall context contains next to the *local context* also the *remote context* of foreign (but locally reachable) mobile systems. As a result of the generic model, the remote context can contain arbitrary domains and entities,

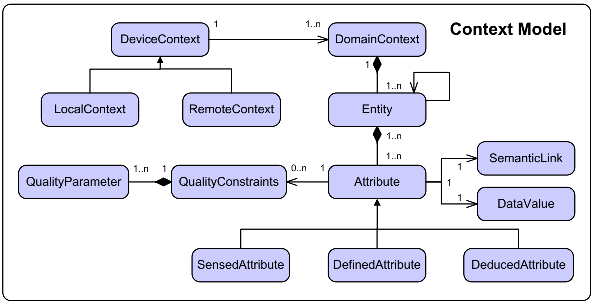


Fig. 5. Schematic Diagram of the Context Model

unaffected by the (limited) set of contextual information modelled within the local context.

3.2 Context Management System

In order to achieve the required separation between context model and management, all middleware functions concerning the management of the presented context model are realized by autonomous services: The central component of the context management is the *RemoteContextService* which manages the exchange of context information between cooperating devices. It implements a protocol to provide local context to other devices and to integrate remote context to learn about the environment of foreign participants. Additionally, a *SecurityPrivacyService* determines which information should be revealed to other participants. A *SerializationService* is responsible for serialization and deserialization of context information by, e.g. representing the model in XML format or others. If context data is required to have a different representation, the *TransformationService* supports the conversion of contextual data values to equivalent formats. Refreshing data values of the context model, which is especially relevant for dynamically sensed attributes, is done by the *UpdateService* which obtains recent raw data from sensors and external services. A *NotificationService* observes the local context model and the relevant context of remote systems in order to allow proactive actions through event notifications. Finally, a *StorageService* saves context information in order to analyse past values and to make predictions about future contexts.

4 Evaluation

In order to evaluate the applicability of the generic context model and management system as proposed above, a prototype implementation of all presented components has been realized within the existing DEMAC middleware which provides basic support for the exchange and distributed execution of mobile processes (cp. [8]) in heterogeneous open system environements. For testing the integrated approach, a distributed example application has been implemented based on a set of mobile processes, each modelling a user-centric workflow to execute a sequence of manual tasks and automatic services. The mobile processes have been modelled with the *DEMAC process description language (DPDL)* (cp. [9]) and contain several activities whose execution is constrained by different contextual criteria, depending on the user's individual requirements as well as on the capabilities of the environment the process is migrating to.

The prototype implementation of the DEMAC middleware runs on different mobile devices, such as a notebook or a PDA as well as on Desktop PCs. Each collaborating device holds its own *local context model* and can obtain relevant *remote context models* if needed. Each local context includes a *DefaultDomain* with a *DeviceEntity* and either a *UserEntity* or a *CommunicationEntity* and a couple of basic context attributes, like static ones (e.g. user identity, display resolution and operating system) and dynamic attributes (e.g. available network connections, free disk space and current location). Furthermore, different services (either locally or remotely accessible) have been realised to execute upcoming (sub-) tasks of the mobile process.

Since context attributes can be exchanged and used during runtime, the processes' propagation can be controlled depending on relevant context data. This hinders mobile process to be transferred to devices outside the right context - thus avoiding unnecessary migrations. For example, in the prototype implementation, the middleware's context service is able to consider the availability of a specific participant (user or device), appropriate services to perform an upcoming task, or the compliance to non-functional requirements like network quality or memory resources. In case the context service cannot find appropriate participants with an acceptable context, the mobile process can be transferred to an arbitrary participant in order to enter a new execution environment and, thus, new contexts.

To verify the results obtained by the analytical model (c.p. section 1.1) and to analyze the applicability of the context service, the evaluation includes an experiment to determine whether or not the probability of successful execution increases by context-based cooperation based on the developed context model. The environment for the experimental series consists of a simple process with one single activity, six heterogeneous devices (as described above) with two devices having the capability to execute the processes' activity and four devices being unable to do so. Because sender and receiver of the mobile process must not be the same, there are 5 possibilities for each process to migrate from one device to another, which makes an execution probability of p=40% within the entire system. To test the behaviour of the prototype under load, three test runs were carried out, each including a number of 100 processes. Figure 6 shows the average number of hops resulting from migration necessary to execute the process successfully. The analysis of the experiments further shows that only a few hops suffice to increase the probability of successful execution to levels more than twice as high. The probability estimated in the analytical model and the applicability of the context model can therefore be confirmed also by practical experimentation.

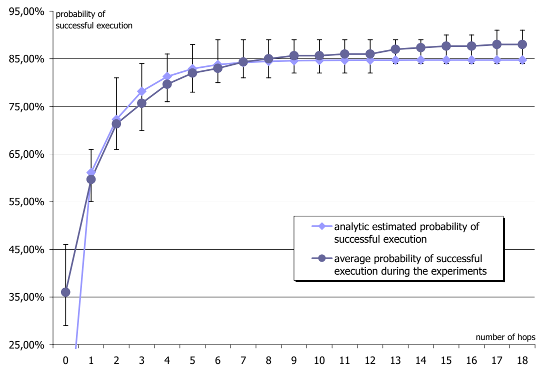


Fig. 6. Experimental Series: Average Probability of Successful Execution

Nevertheless, a known issue of exchanging context data when no suitable participant can be found is the low performance of retrieving remote context information due to the necessity to collect it in recurrent intervals. Besides, better process routing could be achieved, if context data could be shared transitively among participating devices. Furthermore, not all relevant security issues could be resolved so far.

5 Conclusion and Future Work

This paper argues that *mobile processes* can be used to realize a class of context-based systems which use context information to support distributed cooperation by joint use of all combined capabilities of all different (mobile) devices in a distributed environment. Since established context models and management systems do not suffice to support such applications, requirements for middleware support for such a class of *context-based cooperation* were identified. A unified approach of an *adjusted and contemporary context model* and corresponding *management component for mobile systems* were presented and, also based on a respective prototype implementation, evaluated. Future work includes solutions for still open security issues as well as the development of advanced strategies to increase migration performance.

References

1. Bardram, J.E.: The Java Context Awareness Framework (JCAF) - A Service Infrastructure and Programming Framework for Context-Aware Applications. In: Gellersen, H.-W., Want, R., Schmidt, A. (eds.) PERVASIVE 2005. LNCS, vol. 3468, pp. 98–115. Springer, Heidelberg (2005)
2. Baus, J., Krüger, A., Wahlster, W.: A resource-adaptive mobile navigation system. In: IUI 2002: Proceedings of the 7th international conference on Intelligent user interfaces, pp. 15–22 (2002)

3. Chen, H., Finin, T., Joshi, A.: An Intelligent Broker for Context-Aware Systems. In: Dey, A.K., Schmidt, A., McCarthy, J.F. (eds.) UbiComp 2003. LNCS, vol. 2864, pp. 183–184. Springer, Heidelberg (2003)
4. Dürr, F., et al.: Nexus – A Platform for Context-Aware Applications. In: 1. Fachgespräch Ortsbezogene Anwendungen und Dienste der GI-Fachgruppe KuVS (2004)
5. Chen, G., Kotz, D.: Solar: A pervasive computing infrastructure for context-aware mobile applications. Technical report (2002)
6. Held, A., Buchholz, S., Schill, A.: Modeling of Context Information for Pervasive Computing Applications. In: Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2002) (2002)
7. Indulska, J., Robinson, R., Rakotonirainy, A., Henricksen, K.: Experiences in Using CC/PP in Context-Aware Systems. In: Chen, M.-S., Chrysanthis, P.K., Sloman, M., Zaslavsky, A. (eds.) MDM 2003. LNCS, vol. 2574, pp. 247–261. Springer, Heidelberg (2003)
8. Kunze, C.P., Zaplata, S., Lamersdorf, W.: Mobile Process Description and Execution. In: Eliassen, F., Montresor, A. (eds.) DAIS 2006. LNCS, vol. 4025, pp. 32–47. Springer, Heidelberg (2006)
9. Kunze, C.P., Zaplata, S., Lamersdorf, W.: Mobile processes: Enhancing cooperation in distributed mobile environments. Journal of Computers 2(1), 1–11 (2007)
10. McCarthy, J., Buvač, S.: Formalizing Context (Expanded Notes). In: Computing Natural Language. CSLI, vol. 81, pp. 13–50. Stanford University (1998)
11. Salber, D., Dey, A.K., Abowd, G.D.: The Context Toolkit: Aiding the Development of Context-Enabled Applications. In: Proceedings of CHI 1999 (1999)
12. Schilit, B.N., Adams, N., Want, R.: Context-Aware Computing Applications. In: Proceedings of the 1^{st} International Workshop on Mobile Computing Systems and Applications, pp. 85–90 (1994)
13. Strang, T., Linnhoff-Popien, C.: A Context Modeling Survey. In: Davies, N., Mynatt, E.D., Siio, I. (eds.) UbiComp 2004. LNCS, vol. 3205, Springer, Heidelberg (2004)
14. Hofer, T., et al.: Context-Awareness on Mobile Devices - the Hydrogen Approach. In: HICSS 2003: Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS 2003) - Track 9, pp. 292–302 (2003)
15. Wang, X.H., Zhang, D.Q., Gu, T., Pung, H.K.: Ontology Based Context Modeling and Reasoning using OWL. In: PERCOMW 2004: Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications Workshops, p. 18 (2004)
16. Weiser, M.: The Computer for the Twenty-First Century. Scientific American 256(3), 94–104 (1991)

Hydra – An Application Framework for the Development of Context-Aware Mobile Services

Thorsten Caus, Stefan Christmann, and Svenja Hagenhoff

Department of Information Systems and E-Business, University of Göttingen, Platz der Göttinger Sieben 5, 37073 Göttingen, Germany
{tcaus,schrist,shagenh}@uni-goettingen.de

Abstract. A lack of standardisation concerning the handling of context-information and high technical complexity involved in the development of mobile applications are reasons for their low market penetration so far. This article presents 'Hydra', a framework to simplify and shorten the development of context-aware mobile applications and therefore to make the development of these applications more economic. The objective of this paper is to give an idea of how to support developers creating context-adaptive mobile services. An implementation shown in this paper demonstrates the usability of this concept for XHTML-based information services.

Keywords: Context-adaptation, mobile services, mobile applications, framework.

1 Introduction

Broadband mobile networks and increasingly more powerful end devices promise a high potential for universally-accessible mobile data communication applications [1] abbreviated 'mobile applications' in the following. Next to high data transfer costs and missing user acceptance [2][3], the high complexity and lack of standardisation in the development of mobile applications have emerged as the main reasons for their minimal proliferation, and can be indentified as a research gap [4][5].

This article is to introduce a framework that reduces this research gap. It enables that often-used and complex functions of mobile applications (context adaptation, localisation and mobile payment) are encapsulated and, therefore, can be reused. The goal during the design of the framework is to reduce the development effort and the development costs for context-adaptive mobile applications, and to encourage the proliferation of these applications.

This article begins with the motivation to develop the Hydra framework in Section 2. In Section 3, the requirements and subtasks of a context-adaptive application to be supported will be specified. In the subsequent section, the basic construction of the framework will be described from a technical standpoint. In Section 5, two example implementations of Hydra-framework-based services for context-based adaptation of contents will be presented. Finally, Section 6 provides an outlook on further developmental steps of the framework, followed by the conclusion.

W. Abramowicz and D. Fensel (Eds.): BIS 2008, LNBIP 7, pp. 471–481, 2008.

2 Motivation

Because of their compact construction, mobile devices like mobile telephones, PDAs and pagers exhibit structural limitations, especially in the area of multimedia content processing. Small screens, little information processing capacities and uncomfortable navigation and data entry are among these restrictions [6][7].

One possibility to reduce the problem of representing and selecting content in mobile internet use is to only offer users content which is relevant to their particular situation, through context-aware information processing. Thus, part of the scope of the offer is purposely factored out [8].

Which information is relevant to the user in a specific situation can depend on the technical (e.g. bandwidth and available resources), social (e.g. personal characteristics and preferences) or physical context (e.g. location, temperature and time) of the situation [9]. Personalisation (understood here as the adaptation of information to the social context) and localisation (adaptation to the location of the user), in particular, should be considered during information processing. Through this adaptation to the context of the user, mobile applications become more attractive, as they are more readily usable. Mobile phone carriers, in particular, are interested in substantial use of these services, given that more substantial internet use is usually connected with the use of mobile applications. This offers the opportunity to increase average revenue per user. Experts see the future of mobile communication in this context-aware content provision [10].

It has been attempted for a long time to guarantee user-friendly access to mobile services through context-aware information provision. A certain plurality of academic concepts concerning this dilemma currently exists; these, however, find only restricted use in practice. Most examples of context-aware applications represent isolated solutions which are based on a ‘closed-world-assumption’ [11]. Especially from a commercial standpoint, these systems are of only minimal importance [4].

The following reasons for the failure of expectations with respect to the proliferation and use of context-specific mobile applications are discussed in the literature: *(1)* Many singular, specific technologies exist, so that neither provider nor user has any amount of security with respect to the upward compatibility of chosen solutions. For example, up to this point, no consistent standard for the compilation of user profiles has established itself within the domain of personalisation [12][6]. *(2)* The development of these processes is quite complex, which has a negative effect on the pricing and the market launch date [13]. *(3)* Context-specific applications up till now (e.g. restaurant finders, convention guides) are, as a rule, specifically tailored to an application field, and, therefore, are not transferable or adaptable to other problems [6].

Up to this point, individual solutions haven’t exhibited any standardised user interface, nor do they offer trustworthy and comfortable processing of personal data. This can have a negative effect on the acceptance and usability of these applications.

On the other hand, functions such as the processing of contexts have been developed repeatedly in these individual solutions. However, in general, they are not compatible with one another. Thus, development becomes cost-intensive, which reduces willingness to offer mobile services.

Table 1. Common attributes of mobile applications

Attribute	Occurence				
Context-adaption (geographical reference)	Manual	RFID	Cell based	GPS	-
Context-adaption (preference reference)	Manual procedure		Automatic procedure		
Mobile device	Mobile phone/ Smartphone	PDA / MDA	Notebook	Special Device	
Client type	Thin-Client	Fat-Client (e.g. J2ME)	Fat-Client with local data management		
Communication technology	GSM / GPRS / EDGE	UMTS / HSDPA	SMS / MMS	WLAN	Bluetooth

In summary, it can be held that the development of context-adaptive applications is quite complex, and that providers as well as users are confronted with a certain heterogeneity of approaches. Table 1 shows a morphological box [14] which illustrates this heterogeneity through different characteristics of mobile applications.

All of the specifications shown in the morphological box can be found in research or practice; those marked in grey occur relatively often [14].

Two concepts for context adoption can be mentioned in this context: (1) JCAF – Java Context Awareness Framework - a framework to support the development of context-adaptive applications (JCAF, Bardram 2005), which does not focus on mobile applications. (2) Mobilife - an architecture for context-adaptive mobile services which was developed within the scope of the EU project Mobilife (Klemettinen 2007); the focus in this case was on the management of context information. It was about how to personalize the presentation of content, but does not focus on the personalization of content itself.

Up till now, there wasn't any perfected conception for the development of context-adaptive mobile applications which addresses the above-mentioned problem areas and supports the manifestations of mobile applications represented in Figure 1. This research gap is to be reduced by means of the framework for context-adaptive applications represented in this paper.

3 Requirements on a Framework for Context-Adaptive Mobile Applications

3.1 Functional Requirements

As a central barrier to the proliferation and use of context-specific mobile applications, the complexity affiliated with the realisation of such applications was mentioned in the previous remarks. Therefore, it should be deliberated how the developers of such applications can be supported in overcoming this complexity. One obvious solution is to implement complex functionalities once and then reuse them in various applications. This elementary approach corresponds to the encapsulation known from object orientation: Reusable functionalities are encapsulated in classes or class libraries and are accessible through well-defined interfaces. In this way, the complexity of these functionalities is hidden from the developer using them. An encapsulation of functionalities is especially reasonable when the functionalities are sufficiently complex and needed in various applications. Within the framework, the following two areas are supposed to be realised:

Context-adaptation: The application should consider the preferences and the location of the user when processing information. Therefore, personalisation functions (administration of user profiles, transmission of personal data to services and applications with consideration of data protection concerns) must be made available. No solution for the development of mobile applications has been established up to this point (cf. Section 2). On the other hand, localisation functions should be offered in encapsulated form.

Radio technology: In order to provide the mobile user with information independent of the available technology, communication processes (cf. Figure 1) must be supported for conventional wireless technologies (GPRS, UMTS, WLAN, Bluetooth).

3.2 Architectural Requirements

Considering the fact that the most varied end devices and different client architectures must be supported, requirements for the architecture of the framework are as follows:

End devices: An architecture that is executable on the most varied hardware platforms, and that supports, in particular, mobile phones, smart phones and PDAs, is needed. Notebooks and other special devices which, with respect to size and application, do not correspond to the requirements for a constant mobile companion (and, additionally, must be booted) are not to be examined.

Client types: Thin client applications (by which the display of information occurs in XHTML format), as well as fat client applications (which require further program logic on the end device, e.g. in order to store application data) must be supported. To this end, an application manager should make possible the administration, publishing and downloading of further program logic in the framework.

4 Design of the Hydra-Framework

4.1 The Hydra-Framework as a Middleware between Applications and Technologies

In order to facilitate the administration of framework-based applications and to increase the expandability of the framework, the framework, applications and technologies are organized into a conventional layered structure. The framework, therefore, represents a layer between radio technologies and concrete applications, with the goal of reducing the complexity mentioned in Section 3. Thus, the framework functions as a middleware and facilitates the development of applications. In the following this architecture is described and it is also introduced how applications and services can be set up with it.

There are three different layers which must be briefly elucidated (cf. Figure 1): The area *Mobile Communication (Layer 1)* must create the technological basis at the lowest architectural level. It gives the framework above it the ability to automatically execute communication processes in order to exchange information with other framework components through different mobile technologies (e.g. GSM, UMTS, Bluetooth or WLAN). Moreover, this layer contains methods to stream multimedia information between client and server.

The *Framework component (Layout 2)* represents the middle level of the architecture and therefore the connection between the base technologies (Layer 1) and the concrete applications (Layer 3). The framework offers a frame to implement different personalised applications. It contains fundamental mechanisms for personalisation and localisation, as well as methods for the transmission and management of personal data.

Mobile Applications (Layer 3) can be developed which, by means of the framework, have simplified access to the functionalities described above.

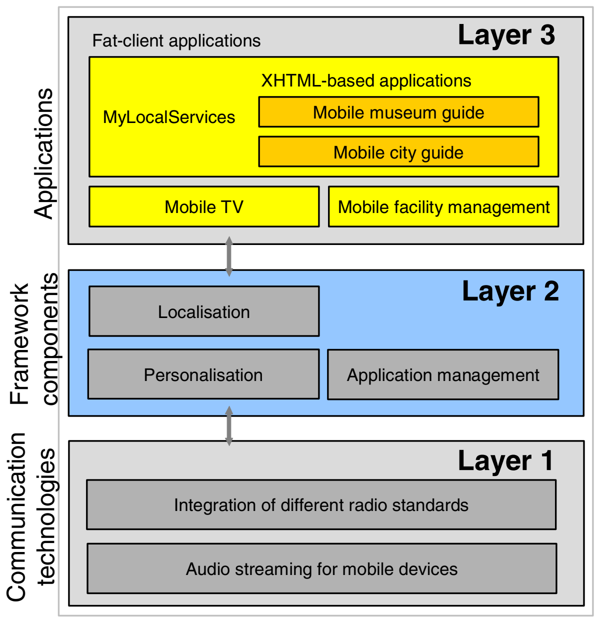


Fig. 1. Hydra as a middleware

4.2 Support of Different Client Types

With consideration of the application types to be supported, two basic application areas are to be examined more closely within the scope of the development of this framework: (1) Thin-client applications, which do not need any further program logic on the mobile device (besides the framework) and utilise server-side personalisation and display information in XHTML. (2) Fat-client applications, which need additional program logic on the client side.

Table 2. Supported types of applications

Application type	Sample applications	Function of the framework-client	Function of the framework-server	Presentation of content
XHTML -based application	- Mobile museum guide - Mobile city guide - Mobile marketing	- Profile management - Presentation of XHTML-contents	- Personalisation of XHTML-contents with respect to user profiles and the location of the user	XHTML
Fat-client	- Mobile TV - Mobile systems to support public transport - Mobile facility management	- Profile management - Presentation and storage of contents using Java - Client personalisation feasable using additional application logic	- Framework-server is optional - Personalisation of XHTML- or different types of content with respect to user profiles and the location of the user	Preparation by Java-application or XHTML

Table 2 provides an overview of the essential particulars of the two application types. In order to let XHTML-based applications be shown on the end devices, the fat-client application *MyLocalServices* was developed on the basis of the framework. XHTML-based thin-client applications can thus be understood as services of the application *MyLocalServices*.

Different context-adaptive services offering XHTML-based information can be supported (e.g. personalisable museum guides or personalisable city guide, cf. Section 5). A client-adjusted server component is necessary to support thin-client applications, as the client does not contain any service logic. This program logic is located at a framework server that can access the context information of the client and is able to prepare content as XHTML-data and to distribute it to the client.

4.3 Realisation of the Context-Adaptation

In order to realise a preparation and provision of information according to the needs of the user, we use the preferences of the user in the form of user profiles (social context) and, on the other hand, the location of the user (physical context). Because of the manifold application types supported by the framework, a partition was conducted according to the client-server principle.

In order to process customer-oriented data, storage of user profiles exclusively on the user end device is used. A mechanism that informs the user what information the service requires is necessary to store the user profile on the mobile device. The framework uses a suitable profile negotiation process, one which avoids multiple unnecessary inputs by the user in that it can access stored generic personalisation information (e.g. age, sex) for the adaptation of new services and applications. In order to achieve this, a component which can create and edit service-specific profiles was implemented.

Framework server: With the framework server, context-adaptive applications can be developed, managed and distributed. Server and client components are coordinated so that they can initiate communication processes independently during the use and personalisation of applications (cf. section 4.4). The application developer can access context information in a standardised fashion by means of the framework server and use these to initiate the adaptation of the application.

Framework client: The framework client encapsulates the processing of user profiles, as elucidated in Section 4.3. In addition, terminal and cell-based localisation are made possible. A direct positioning by means of the Global Positioning System (GPS) or an indirect positioning by means of a cell-based localisation are the most conceivable in this case.

During this process, the geocoordinates of the next GSM base station can be ascertained via proximity sensing. If the framework client has connected to a framework server through short-range radio technology (e.g. Bluetooth), clients can do without additional devices like GPS receivers or the similarly-unprecise localisation by means of the GSM-mobile network, as the location of the server components can be seen as the location of the client.

In addition to the simplified process of determining geocoordinates by means of the location of a user, the address of the current position can be determined automatically. An interface with the geoinformation system MapPoint from Microsoft was developed for this purpose. However, it is also possible to use alternative services, like, for example, GeoNames or OpenStreetMap.

4.4 Communication Processes

The communication processes when using the application *MyLocalServices* are exemplified. The communication between client and server components for other applications or radio technologies occurs analogous to this. In this example, the user is located in a museum, uses the short-range radio technology Bluetooth and wants to see all available services with a mobile device for the information system ‚museum'. The following communication processes occur: After the framework has been activated, the mobile device sends a connection request with an empty session template to the information system by means of a Bluetooth hotspot located in range. The receipt of the empty session template signals to the server components that the user was not previously connected with one of their services. Service descriptions therefore are sent to the mobile device and, subsequently, are summarised graphically in an overview.

If the user changes location, so that his mobile device connects with another access point, then it should be able to continue the service without any difficulty. For this reason, the framework stores the current session template if the mobile connection (in this example, Bluetooth) to the end device is interrupted. This session template contains a unique identification of the service used, as well as a session ID, which makes it possible for the service to locate the previously-stored service profile and allows a continuation of the information provision. If a long-range radio technology like GSM, GPRS or UMTS is implemented, sending an empty session template, as well as service descriptions, becomes unnecessary. In this case, the framework supports the exchange of personalisation templates.

In order to guarantee communication between these subsystems, interfaces are to be provided on client- and server-side.

In order to display the XHTML contents (required for the application *MyLocalServices*), a web browser is needed on the user's mobile device. For the development of the framework, a rendering engine to display the XHTML contents was therefore created and integrated into the framework.

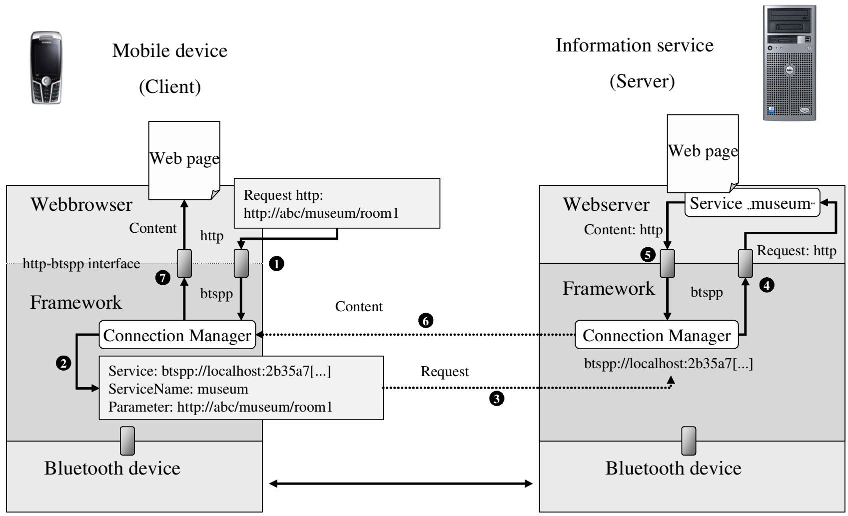


Fig. 2. Client request for service contents (example: mobile museum guide)

The support of Bluetooth as a transmission technology provided as an example requires special treatment during the transport of contents. For this reason, this will be elucidated in greater detail by means of Figure 2: The framework-client transmits the requested HTTP address to the server as a parameter. The server then generates the contents and sends them back to the client over the framework, which routes them to the browser for display. Client and server interfaces between the browser/web server, the framework and the communication technology Bluetooth are necessary for this. Sending web requests and receiving corresponding contents can occur in a simple format, as only the target address in HTTP and a session ID are needed. The communication processes for profile negotiation, however, exhibit a higher complexity. For this reason, these messages are structured by means of XML schemata, which, because of their higher proliferation and acceptance, are supported extensively.

4.5 Application Manager

The realisation of the application manager turns out to be complex: J2ME, which has been used for implementation because of end device independency, does not allow adding program logic using class loading [15]. Another possible option is to add the framework as a package for each application and then install the concrete applications as J2ME-program-suites. This leads to the critical disadvantage that on the one hand the framework is arranged several times on the mobile device (for each application separately). On the other hand, critical personal data has to be stored on an external memory card because J2ME applications in principle do not have access to the memory of other J2ME applications.

In this way an overall management of person data and a trustful treatment of appropriate information are not possible. For that reason, an application manager has been developed and integrated in the framework. It allows searching for, paying and installing further program logic. When installing further applications, the client sends the desired program and a list of all installed framework-applications to a central framework-server. The central server builds an installable mobile Java-application, which contains the framework, the already installed and the desired application.

The client installs this package in form of a program-update under J2ME. Personal data and program options stored before remain untouched. As the application manager needs access to all framework-applications, applications have to be managed and stored centrally.

5 Evaluation

In order to demonstrate the functional efficiency of this framework, two services for the *MyLocalServices* application were developed. A *personalisable museum guide* which communicates over Bluetooth and allows a personalised museum tour dependent on the location (exhibition room where the Bluetooth transmitter has been installed) and the characteristics and preferences of the user (language, level of detail, age and gender) has been implemented. An extension of the preference profile to art eras, favoured style and artist based on the framework is easily achievable. The service provided informs the visitor with pictures and text of the museum area he has just entered. Additionally, a *personalisable city guide* was implemented. In this case, a server unit in a public-transit bus communicates via Bluetooth with the mobile telephones of the passengers and considers their preferences (language, sightseeing categories), as well as their location (framework server in the bus which localises itself via GPS) in order to be able to present interesting sights to the passengers in time.

Figure 3 shows screenshots of the use of the city guide service example, which was created based on the framework and the application *MyLocalServices*. The exact information that should be accessed for personalisation can be defined in a personalisation template in XML (cf. Section 4.3). Figure 4 shows the proper template for the mobile city guide. This profile is automatically processed (cf. figure 3 bottom center), stored and used again on the client for the next connection to the city guide. After finishing the profile transfer, the framework server forwards the request to the in each case dedicated service class.

Using the framework, the development of context-adaptive applications becomes much easier for programmers. They need only to create a service class based on a programming interface provided by the framework and to store this within the service directory of the framework server. Subsequently, the framework server registers this new service and calls up the method *doGet(…)* within the service class on request. At the same time, the framework server delivers an object of the class *profile*, which contains the XML user profile, as well as methods to effectively search this. Additionally, the framework server can localise itself if the proper hardware is connected via Bluetooth or USB.

Fig. 3. Context-adaptive mobile city guide running on framework

```
<?xml version="1.0"?>
<serviceProfile>
  <choiceGroup>
    <name ref="language">Language</name>
    <choice ref="german">German</choice>
    <choice ref="english">English</choice>
  </choiceGroup>

  <choiceGroup>
    <name ref="interests">Preference list</name>
    <choice ref="culture">Culture</choice>
    <choice ref="important facilities">
        Important facilities</choice>
    <choice ref="restaurants">Restaurants</choice>
    <choice ref="disco">Discotheques</choice>
    <choice ref="bars">Pubs und bars</choice>
    <choice ref="shopping">Shopping</choice>
  </choiceGroup>
</serviceProfile>
```

Fig. 4. Personalisation template for the mobile city guide

6 Conclusions and Future Work

In this text, a framework for the provision of context-adaptive, mobile applications was introduced, one which can standardise, simplify and make more cost-efficient the implementation of these applications. Based on the prototypical partial implementtation of the concept and the example application developed, the principle practicality of this framework for fat client applications (*MyLocalServices*), as well as for thin client applications (personalisable city guide) could be demonstrated. Based on this, manifold services could be more simply offered.

Beside the advantages of the framework (standardisation, encapsulation of communication, personalisation and standardised interface), certain obstacles which could impede the implementation of this technology must be taken into account. Firstly, the framework must possess a certain awareness level in order to be considered and implemented by developers. Thus it is afflicted by the standard problems of network effect goods.

Therefore, future work will, on the one hand, concentrate on examining the already-existing prototypical framework as it pertains to user acceptance (user study on mobile city tours) and commercialising services of the framework (e.g. the mobile city guide), and, on the other hand, on supplementing the framework with functionalities which have not yet been realised (e.g. mobile payment).

References

1. Hess, T., Figge, S., Hanekop, H., Hochstatter, I., Hogrefe, D., Kaspar, C., Rauscher, B., Richter, M., Riedel, A., Zibull, M.: Technische Möglichkeiten und Akzeptanz mobiler Anwendungen – Eine interdisziplinäre Betrachtung. Wirtschaftsinformatik 47/1, 6–16 (2005)

2. Gerum, E., Sjurts, I., Stieglitz, N.: Der Mobilfunkmarkt im Umbruch, DUV, Wiesbaden (2003)
3. Anckar, B., D'Incau, D.: Value-added services in mobile commerce: an analytical framework and empirical findings from a national consumer survey, System Sciences. In: Proceedings of the 35th Annual Hawaii International Conference, HICSS (2003)
4. Kaspar, C., Diekmann, T., Hagenhoff, S.: Kontextadaptive Systeme. In: Wirtschaftswissenschaftliches Studium, 9/2005 (2005)
5. Acampora, G., Loia, V., Nappi, M., Ricciardi, S.: Ambient intelligence framework for context aware adaptive applications. In: Computer Architecture for Machine Perception, CAMP (2005)
6. Billsus, D., Pazzani, M., Chen, J.: A Learning Agent for Wireless News Access. In: Proceedings of the 5th international conference on Intelligent user interfaces, pp. 33–36 (2000)
7. Smyth, B., Cotter, P.: Intelligent Navigation for Mobile Internet Portals. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI 2003), Acapulco (2003)
8. Diekmann, T., Kaspar, C., Seidenfaden, L., Hagenhoff, S.: Kontextbewusste Informationsdienste auf Grundlage von Information Beacons. In: Lehner, F. (ed.) Multikonferenz Wirtschaftsinformatik, Passau (2006)
9. Schilit, B., Adams, N., Want, R.: Context-Aware Computing Applications. In: IEEE Workshop on Mobile Computing Systems and Applications (1994)
10. Magedanz, T.: Communications Platforms for Service Orchestration, Conference on Growth through Mobile Communications, Münchner Kreis, Munich (2006)
11. Chen, G., Kotz, D.: A survey of Context-Aware Mobile Computing Research. Dartmouth Computer Science Technical Report TR2000-381, Accessed 2006-10-10 (2006), http://www.cs.dartmouth.edu/~dfk/papers/chen:survey-tr.pdf
12. Wörndl, W.: Privatheit bei dezentraler Verwaltung von Benutzerprofilen, Technische Universität München, Munich (2003)
13. Dey, A., Futakawa, M., Salber, D., Abowd., G.: The Conference Assistant: Combining Context-Awareness with Wearable Computing. In: Proceedings of the 3rd International Symposium on Wearable Computers, ISWC 1999, San Francisco, pp. 21–28 (1999)
14. Fraunhofer Institut Software und Systemtechnik: Mit Explore Museen neu erleben und entdecken, Accessed 2006-09-15 (2006), http://www.do.isst.fraunhofer.de/deutsch/download/38855_PB_EXPLORE_0904_de.pdf
15. Schild, T.: Ein Überblick über die technologischen Möglichkeiten und Erfahrungen aus der Praxis, Zühlke Engineering, Frankfurt (2002)

Author Index