



eKNOW 2015

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

ISBN: 978-1-61208-386-5

February 22 - 27, 2015

Lisbon, Portugal

eKNOW 2015 Editors

Dirk Malzahn, Dirk Malzahn Ltd / HfH University, Germany

Conceição Granja, Norwegian Centre for Integrated Care and Telemedicine,
University Hospital of North Norway, Norway

eKNOW 2015

Forward

The seventh edition of the International Conference on Information, Process, and Knowledge Management (eKNOW 2015) was held in Lisbon, Portugal, February 22 - 27, 2015. The event was driven by the complexity of the current systems, the diversity of the data, and the challenges for mental representation and understanding of environmental structure and behavior.

Capturing, representing, and manipulating knowledge was and still is a fascinating and extremely useful challenge from both theoretical and practical perspective. Using validated knowledge for information and process management and for decision support mechanisms raised a series of questions the eKNOW 2015 conference was aimed at.

eKNOW 2015 provided a forum where researchers were able to present recent research results and new research problems and directions related to them. The topics covered aspects from knowledge fundamentals to more specialized topics such as process analysis and modeling, management systems, semantics processing and ontology.

We take this opportunity to thank all the members of the eKNOW 2015 Technical Program Committee as well as the numerous reviewers. The creation of such a broad and high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to the eKNOW 2015. We truly believe that, thanks to all these efforts, the final conference program consists of top quality contributions.

This event could also not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the eKNOW 2015 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that eKNOW 2015 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in knowledge management research.

We also hope that Lisbon provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city.

eKNOW 2015 Chairs

Dirk Malzahn, OrgaTech GmbH, Germany

Roy Oberhauser, Aalen University, Germany

Conceição Granja, Norwegian Centre for Integrated Care and Telemedicine, University Hospital of North Norway, Norway

eKNOW Special Area Chairs

Technological foresight and socio-economic evolution modelling

Andrzej M.J. Skulimowski, AGH University of Science and Technology - Krakow, Poland

eKNOW 2015

COMMITTEE

eKNOW Advisory Chairs

Dirk Malzahn, Dirk Malzahn Ltd / HfH University, Germany

Roy Oberhauser, Aalen University, Germany

Conceição Granja, Norwegian Centre for Integrated Care and Telemedicine, University Hospital of North Norway, Norway

eKNOW Special Area Chairs

Technological foresight and socio-economic evolution modelling

Andrzej M.J. Skulimowski, AGH University of Science and Technology - Krakow, Poland

eKNOW 2015 Technical Program Committee

Gil Ad Ariely, Interdisciplinary Center Herzliya (IDC), Israel

Werner Aigner, Institute for Application Oriented Knowledge Processing – FAW / University of Linz, Austria

Panos Alexopoulos, iSOCO, Spain

Jesus Manuel Almendros Jimenez, Universidad de Almería, Spain

Amin Anjomshoaa, Vienna University of Technology, Austria

Zbigniew Banaszak, Warsaw University of Technology, Poland

Ladjel Bellatreche, LISI- ENSMA/ Poitiers University, France

Peter Bellström, Karlstad University, Sweden

Jorge Bernardino, Polytechnic Institute of Coimbra, Portugal

Yaxin Bi, University of Ulster - Jordanstown, UK

Marco Bianchi, Fondazione Ugo Bordoni, Italy

Grzegorz Bocewicz, Koszalin University of Technology, Poland

Sabine Bruaux, Picardie Jules Verne University, France

Elżbieta Bukowska, Poznan University of Economics, Poland

Martine Cadot, University of Nancy1, France

Massimiliano Caramia, University of Rome "Tor Vergata", Italy

Yu Cheng, IBM TJ Watston Research Center, USA

Chi-Hung Chi, CSIRO, Australia

Dickson K.W. Chiu, Dickson Computer Systems, Hong Kong

Paolo Cintia, University of Pisa, Italy

Marco Cococcioni, University of Pisa, Italy

Ting Deng, Beihang University, China

Ioan Despi, University of New England, Australia

Ali Eydgahi, Eastern Michigan University, USA

Francesca Fallucchi, Guglielmo Marconi University, Italy

Abed Alhakim Freihat, University of Trento, Italy

Elvis Fusco, Centro Universitário Eurípides de Marília – UNIVEM, Brazil

Susan Gauch, University of Arkansas, USA

Conceição Granja, Norwegian Centre for Integrated Care and Telemedicine, University Hospital of North Norway, Norway
Gregory Grefenstette, Exalead, France
Pierre Hadaya, ESG UQAM, Canada
Georges Hebrail, Electricité De France (EDF) R&D, France
Juergen Hoenigl, Johannes Kepler University, Austria
Daniela Hossu, University 'Politehnica' of Bucharest, Romania
Vana Kalogeraki, Athens University of Economics and Business, Greece
Khaled Khelif, EADS- Val de Reuil, France
Daniel Kimmig, Karlsruhe Institute of Technology (KIT), Germany
Marite Kirikova, Riga Technical University, Latvia
Frank Klawonn, Ostfalia University of Applied Sciences, Germany
Tomomi Kobayashi, Waseda University, Japan
Andrew Kusiak, The University of Iowa, USA
Franz Lehner, University of Passau, Germany
Johannes Leveling, CNGL, Ireland
Chee-Peng Lim, Deakin University, Australia
Matthias Loskyll, German Research Center for Artificial Intelligence (DFKI), Germany
Dickson Lukose, MIMOS-Berhad, Malaysia
Dirk Malzahn, Dirk Malzahn Ltd / HfH University, Germany
Philippe Marchildon, Université du Québec à Montréal, Canada
Mohammed Ameen Marghiani, King Abdulaziz University, Saudi Arabia
Luis Martínez López, University of Jaén, Spain
Marco Mevius, HTWG Konstanz, Germany
Toshiro Minami, Kyushu Institute of Information Sciences, Japan
Anirban Mondal, University of Tokyo, Japan
Yasuhiko Morimoto, Hiroshima University, Japan
Mirco Nanni, ISTI-CNR, Italy
Roy Oberhauser, Aalen University, Germany
Olasunkanmi Olajide, Federal University of Agriculture, Nigeria
Daniel O'Leary, University of Southern California, USA
Jonice Oliveira, Federal University of Rio de Janeiro (UFRJ), Brazil
Joanna Isabelle Olszewska, University of Gloucestershire, United Kingdom
Sethuraman Panchanathan, Arizona State University, USA
Andreas Papasalouros, University of the Aegean - Samos, Greece
Ludmila Penicina, Riga Technical University, Latvia
Tuan D. Pham, The University of Aizu - Aizu-Wakamatsu, Japan
Lukas Pichl, International Christian University, Japan
Przemysław Pukocz, P&B Foundation / AGH University of Science and Technology, Poland
Lukasz Radlinski, West Pomeranian University of Technology, Poland
P.Krishna Reddy, International Institute of Information Technology Hyderabad (IIITH), India
Ulrich Reimer, University of Applied Sciences St. Gallen, Switzerland
Pierre N. Robillard, Polytechnique Montréal, Canada
Fariba Sadri, Imperial College of Science, Technology and Medicine, UK
Aitouche Samia, University Hadj Lakhdar Batna, Algeria
Jagannathan Sarangapani, Missouri University of Science and Technology, USA
Dobrica Savic, International Atomic Energy Agency, Austria
Erwin Schaumlechner, Tiscover GmbH - Hagenberg, Austria

Giovanni Semeraro, University of Bari "Aldo Moro", Italy
Jungpil Shin, University of Aizu, Japan
Andrzej M. Skulimowski, AGH University of Science and Technology, Poland
Lubomir Stancev, Indiana University - Purdue University Fort Wayne, USA
Ryszard Tadeusiewicz, AGH University of Science and Technology, Poland
Masakazu Takahashi, Yamaguchi University, Japan
Carlo Tasso, Università di Udine, Italy
I-Hsien Ting, National University of Kaohsiung, Taiwan
Jan Martijn van der Werf, Utrecht University, Netherlands
Stefanos Vrochidis, Information Technologies Institute, Greece
Da-Wei Wang, Institute of Information Science - Academia Sinica, Taiwan
Haibo Wang, Texas A&M International University, USA
Hongzhi Wang, Harbin Institute of Technology, China
Hans Weigand, Tilburg University, Netherlands
Peter Wiedmann, HTWG Konstanz, Germany
Shengli Wu, University of Ulster - Newtownabbey, Northern Ireland, UK
Takahira Yamaguchi, Keio University, Japan
Mansour Esmaeil Zaei, Panjab University, India

Copyright Information

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

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

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

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

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

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

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

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

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

Table of Contents

Enhanced Stakeholder Socialization using Common Language in Agile BPM: Living business processes models instead of rigid documentations <i>Marco Mevius, Erich Ortner, and Peter Wiedmann</i>	1
A Versioning and Commenting Approach for Enhancing Group Efficiency in Collaborative Web-Based Business Process Modeling Tools <i>Justus Holler</i>	9
An Ontology for Formalizing and Automating the Strategic Planning Process <i>Juan Luis Dalmau-Espert, Faraon Llorens-Largo, and Rafael Molina-Carmona</i>	17
A Usability Evaluation Methodology of Digital Library <i>Luz A. Sanchez-Galvez and Juan M. Fernandez-Luna</i>	23
LOM, a Locally Oriented Metric which Improves Accuracy in Classification Problems <i>Julio Revilla and Evaristo Kahoraho</i>	29
First Steps Towards a Formal Analysis of Law <i>Tom Maarten van Engers and Robert van Doesburg</i>	36
Study on Web Analytics Utilizing User Environment Segmentation in “Business to Business” site. <i>Akiyuki Sekiguchi and Kazuhiko Tsuda</i>	43
Tacit and Explicit Knowledge in Software Development Projects: Towards a Conceptual Framework for Analysis <i>Hanna Dreyer, Martin Wynn, and Robin Bown</i>	49
Knowledge Processes in German Hospitals. First Findings from the Network for Health Services Research Metropolitan Region Bremen-Oldenburg <i>Lars Rolker-Denker, Insa Seeger, and Andreas Hein</i>	53
S-Grouper A Semantic Based System to Semi-Automatic Encode Hospital Activities <i>Roberta Cuel, Andrea Francesconi, Filippo Nardelli, and Giampaolo Armellin</i>	58
ChoreMAP: Extracting And Displaying Visual Database Summaries Tool <i>Cherni Ibtissem, Faiz Sami, Laurini Robert, and Warghi Mariem</i>	60
The Knowledge Reuse in an Industrial Scenario: A Case Study <i>Gianfranco E. Modoni, Enrico G. Caldarola, Walter Terkaj, and Marco Sacco</i>	66
ARPPA: Mining Professional Profiles from LinkedIn Using Association Rules <i>Paula Silva and Wladmir Brandao</i>	72

Towards Improving Students' Attitudes to Lectures and Getting Higher Grades --With Analyzing the Usage of Keywords in Class-Evaluation Questionnaire-- <i>Toshiro Minami and Yoko Ohura</i>	78
Supporting Provenance in Climate Science Research <i>Brett Yasutake, Niko Simonson, Jason Woodring, Nathan Duncan, William Pfeffer, Hazeline Asuncion, Munehiro Fukuda, and Eric Salathe</i>	84
Discount Coupons Dematerialization: a Comprehensive Literature Review <i>Goncalo Paiva Dias, Helder Gomes, Jorge Goncalves, Daniel Magueta, Fabio Marques, Ciro Martins, Mario Rodrigues, and Jorge Araujo</i>	92
Knowledge Intensive Evolutionary Algorithms <i>Francois de Bertrand de Beuvron, Carlos Catania, and Cecilia Zanni-Merk</i>	99
Guidelines for Social Media Mining for Innovation Purposes. Experiences and Recommendations from Literature and Practice <i>Robert Eckhoff, Mark Markus, Markus Lassnig, and Sandra Schon</i>	106
The Application of Machine Learning to Problems in Graph Drawing - A Literature Review <i>Raissa dos Santos Vieira, Hugo Alexandre Dantas do Nascimento, and Wanderson Barcelos da Silva</i>	112
Game Refinement Theory and Multiplayer Games: Case Study Using UNO <i>Alfian Ramadhan, Hiroyuki Iida, and Nur Ulfa Maulidevi</i>	119
ERP Implementation in a Developing World Context: a Case Study of the Waha Oil Company, Libya <i>Hosian Akeel and Martin Wynn</i>	126
The Key Contributions of the Operations Management and Information Systems Disciplines to Business Process Management <i>Philippe Marchildon and Pierre Hadaya</i>	132
Mapping the Fuzzy Semantic Model into Fuzzy Object Relational Database Model <i>Sabrine Jandoubi, Afef Bahri, Salem Chakhar, and Nadia Yacoubi-Ayadi</i>	138
Knowledge and Technology Transfer in the Center for Scientific and Technical Information of the Wroclaw University of Technology <i>Anna Walek and Katarzyna Kozłowska</i>	144
ODINet Online Data Integration NETwork: an Innovative Ontology-based Data Search Engine <i>Stefania Pieroni, Michela Franchini, Sabrina Molinaro, Alessandro Greco, Francesco Pitto, Moreno Toigo, and Luca Caterino</i>	152

If Experience is Worth, How Experts Behave in a Manga Case 159
Satoshi Takahashi, Toru B. Takahashi, Atsushi Yoshikawa, and Takao Terano

Compound Noun Polysemy and Sense Enumeration in WordNet 166
Abed Alhakim Freihat, Biswanath Dutta, and Fausto Giunchiglia

Automatic Diagrammatic Multiple Choice Question Generation from Knowledge Bases 172
Khalil Bouzekri, Qiang Liu, Husam Yasin, Benjamin Min Xian Chu, and Dickson Lukose

Enhanced Stakeholder Socialization using Common Language in Agile BPM

Living business processes models instead of rigid documentations

Marco Mevius
HTWG Konstanz
Constance, Germany
mmevius@htwg-konstanz.de

Erich Ortner
TECHNUM
Constance, Germany
e.ortner@technum.biz

Peter Wiedmann
AXON IVY AG
Munich, Germany
peter.wiedmann@axonivy.com

Abstract- Business Process Management (BPM) and its supporting systems (BPMS) focus on business processes without sufficiently taking all stakeholders into account. Instead, the focus is put on the modeling and execution of business processes. Rigid business process documentations, long update cycles and insufficient understanding of business process models are the results. In this paper, we present how to use common language to simplify and further develop existing methods aiming at “living” business processes instead of rigid documentations. We enhance the innovative agile method BPM(N)^{Easy1.2}, which includes a supporting mobile application called BPM Touch. This combination allows the powerful usage of common language with BPM and enables stakeholders to better “socialize” with one another in the context of business processes and their management. The application of the method is illustrated with the help of a public administration sample. The paper concludes with a summary and outlook on further research.

Keywords-computational society architecture; business process; common language; agile bpm.

I. INTRODUCTION

Business Process Management is becoming more and more an important discipline in organizations. High dependencies and fast moving changes of the business processes and general business conditions are reasons behind this trend [1]. Because of this importance, many BPM approaches and tools have been developed by different researchers and software companies. These BPM approaches can be categorized in traditional and agile approaches. Both categories deliver frameworks which describe how to handle current BPM challenges, e.g., [2][3][4][5]. In general, the approaches cover all phases of a business process lifecycle, starting with modeling business processes, analyzing and automating through monitoring and optimizing them. Besides the BPM approaches, a lot of BPM systems are available. Only on the German market, Fraunhofer [6] investigated over fifty different BPMSs which support BPM experts, e.g., in modeling or automating business processes. Nowadays, these approaches and tools are increasingly used to set up a successful technical BPM environment in companies. Aside from positive effects, such as measuring Key Performance Indicator (KPI), monitoring of the executed business processes [7], and efficient automated workflows [8], this setup also brings

along several challenges. Failing BPM trainings [9] or outdated business process documentations [10] are well-known examples. Furthermore, the interpretation of business process models is affected by the specific knowledge of each person [11]. These problems arise because of complex implementations and long update cycles or approaches, which are for technical BPM experts only. In the end, the involvement of all stakeholders breaks down. This leads to rigid models instead of “living” business processes within a successful socialized BPM philosophy. From the authors’ point of view, a promising option and solution out of this scenario is the application of agile methods and corresponding tools. Agility is defined as a successful balance between flexibility and robustness [12]. Given samples, e.g., out of the software engineering branch show how successful the usage of agile methods can be. Based on the Agile Manifest [13], which describes rules for interacting in an agile manner, methods such as Scrum [14] or Extreme Programming [15] are applied in a lot of software projects to implement new products successfully. The results of this application are, e.g., decreased costs and higher user acceptance. Also in the branch of BPM agile methods are applied more and more. BPM(N)^{Easy1.2} [4] is one example. BPM(N)^{Easy1.2} describes a combination between BPM and the business process modeling language BPMN 2.0 with the ambition of providing a method which makes the traditional phases of BPM [2] – modeling, analyzing, execution, optimization – more agile and easier.

However, there are still challenges which are not solved in a satisfying way. For instance, synchronization and interaction between all stakeholders are fraught with misunderstandings [16]. Interaction describes the collaboration between the stakeholders in general, synchronization the explicit knowledge update of each other c.f. Mevius et al.[17]. Therefore, business processes are often inconsistent or give a biased view on the reality. Another problem is that the high amount of information and needs regarding a business process cannot be captured without expending a lot of time and effort [18]. But, as Mayr [19] mentioned, the aim should be to “more act, less plan” business processes. In addition, Hauser [20] speaks about “five social feelings”, which have to be included for successful business process management, e.g., the feeling of

anger can be a positive or negative driver of action within a business process activity.

To enable agile BPM to counter these challenges, this paper introduces the application of common language as fundament of methods. Furthermore, this paper describes how existing approaches can be improved with the help of common language and enabled for Social Computing. Therefore, common language defines both, the language which is used for a “normal” conversation and the language which is used for an expert talk [21], e.g., discussion about a parking slot approval by the public administration. Furthermore, from the authors’ perspective, Social Computing describes the stakeholder (human) interactions, which are supported by different Information Technology (IT) systems. Referencing to [22], stakeholders use IT to generate new content together, e.g., new business processes.

The paper is structured as follows: Section 2 describes the evolution towards Social Computing and introduces a Social Computing architecture. Section 3 applies this architecture and analyses related work regarding the usage of common language in BPM and other related disciplines. In Section 4, the application of common language in agile business process management is described by means of the project “Smart City Constance”. Section 5 introduces the existing BPM(N)^{Easy1.2} method and demonstrates where and how common language can be applied within an agile method. Furthermore, a tool support is illustrated. Section 6 presents a conclusion and outlook.

II. SOCIAL COMPUTING

Social Computing empowers individual stakeholders independent of their IT skill level [22]. For instance, stakeholders are enabled to share their creativity or expertise with the help of (Web) applications. Fig. 1 displays the different layers of a Computational Society Architecture.

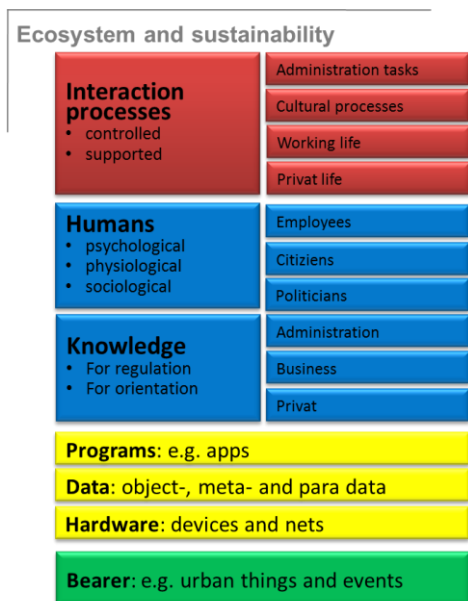


Figure 1. Computational Society Architecture

Furthermore, social interaction and content sharing are becoming more and more important. Pure technique-focused requirements lose in value. First approaches for this shift can already be found in [23]. Ortner [23] describes a matrix, which can be used to capture all requirements for an application system, e.g., business process applications. These requirements can be technical as well as economic or human-oriented.

Based on initial events, IT (c.f. yellow layers in Fig. 1) provides supporting technologies. With the help of these technologies all stakeholders (humans) can interact and synchronize themselves regarding knowledge and the interaction context. Application systems, which are based on this architecture (Fig. 1) depicts how humans, technique and organization work together [24].

The following sections focus on the ecosystem of agile business process management.

III. RELATED WORK

The direct linking of the (human) language with digital applications is a fundamental objective of language-based computer science [21]. If an application system, e.g., a complex workflow application or a customer relationship system, is introduced or developed individually, the requirements of the future end users must be collected and described as completely as possible from a technical point of view [25]. The aim should be the attainment of a mutual understanding of all stakeholders: end users, BPM, IT experts and management.

This section links to existing work selected from BPM (-related) disciplines, in the context of enhancing communication between as well as understanding and involving all stakeholders.

Moody [26] defined principles for designing cognitively effective visual notations, aiming for an optimized human communication and problem solving. For instance, the principle of “Perceptual Discriminability” describes that different symbols should be strictly distinguishable from each other. In [27], Evans introduces the approach of “Domain Driven Design” in which the goal of integrating the user directly in the conception and design phase is claimed. In addition to formulating software requirements, the end users are also involved in the design and modeling phase of the software. On one hand, this increases the understanding of the IT experts for the domain and its issues, on the other hand, end users are introduced in the process early on and are quickly able to detect design errors or misunderstandings. User stories, which are formulated by the end users, serve the purpose of describing the future software system. In general, a user story describes a requirement from the user’s perspective. To develop uniform user stories to different patterns, formulations such as “As a <user role>, I <want/need/can/etc.> <goal> <reason> ” can be used. In this context, communication counts more [28] than the actual software requirement.

Another approach is called Very Lightweight Modelling Language (VLML) [29]. This approach follows the idea of combining the expressive power and ease of natural language, rapid and easy sketching and functions to structure, analyze and validate models. Even these approaches [26][27][29] extend the possibilities, they do not provide an entire support of common language.

Within quality management, a continuous improvement and involvement of stakeholders is required. For this purpose, different approaches, such as total cycle time, Kaizen and Lean Sigma have been developed [30]. For instance, Lean Sigma leads an improvement of the product and process quality, on one hand, and, on the other hand, it increases the performance of business processes. In order to involve stakeholders, audits, which are conducted internally or externally, are used to interview the user directly. As a result of these interviews, the quality management can interact to improve the quality of, e.g., an automated production process. Nevertheless, usually, the audits are too rarely held so that the involvement of the stakeholders are too less.

Schnabel et al. [5] describe how stakeholders can define business process requirements effectively. Hereby, the approach suggests a modeling language (Language for Lightweight Process Modelling (LLPM)), which keeps away (technical) information from the participants, without losing the opportunity of business process automation. Concepts of representation and visual composition are used by Antunes et al. [31]. The approach focuses on the perspective of end users while the modeling of business processes. For instance, business process activities can be enriched by adding annotations in form of text or pictures to a business process element. This leads to an enhanced understanding of the models. Another method [33] suggests the connections of additional information as natural language artifacts to business processes elements. Within this artifact, the identifier label of a business process activity is mentioned and connected. By providing this “common understandable” information, a higher integration of employees, especially of the operative business (non-IT-specialists) can be reached. Furthermore, Bruno et al. [34] introduce a method, which focuses on Social Media integration in BPM. Hereby, they use different aspects, such as real-time collaboration to support the activities of BPM. Rigid structures, which are identified in, e.g., models of business process or role models, can be relaxed. However, especially in agile environments even higher flexibility is required, e.g., not only the feedbacks of all stakeholders have to be collected frequently. Moreover, automated business processes need to be developed and adapted very flexible.

Last, but not least, there are informal options for communicating with stakeholders defined and implemented by industry players [26]. Among these, for instance, existing notation standards as the modeling language Business Process Modeling and Notation (BPMN 2.0) [35] are

modified for a specific goal. One approach comes with a BPM suite called Axon.ivy [36]. Within the tool, the BPMN 2.0 standard elements can be enriched in different ways, e.g., by adding user interface drafts directly on an activity. Furthermore, Barker [37] has defined a natural-language-based normative language which is used by Oracle modeling tools.

All sketched approaches allow a far-reaching usage of BPM. However, there is still a critical gap between stakeholders e.g. end users and IT experts. Moreover, the stakeholder involvement in BPM projects and their success is still not satisfying, c.f. study to BPM projects [18]. During the next sections it will be illustrated how the interaction and synchronization of all stakeholders, especially involving end users, can be enhanced within existing methods by using common language.

IV. SCENARIO





A business process model has been chosen to demonstrate the active usage of common language with agile BPM. The business process is taken from a real business process model repository of the innovative project “Smart City Constance”. The project deals with Social Computing in the context of public administration. The business process model describes the sequence of activities which have to be executed to approve a request for reserved car parking. Furthermore, the involved stakeholders each bring along different knowledge and motivation, e.g., a citizen wants to have a parking space as soon as possible, but the public authority has to follow the predefined business process. The scenario describes two iterations of capturing, modeling the business process and involving all required stakeholders. The business process model is modeled with BPMN^{Easy1.2} and executed on the mobile application BPM Touch [38]. BPMN^{Easy1.2} is a business process modeling language which uses BPMN 2.0, but only with a specific element set. In addition, it is possible to add some media files, e.g., video sequence to the modeled concepts. Table I illustrates an extract of the first iteration of capturing information, which is required for an initial version of a business process model.

TABLE I. EASY CAPTURE SHEET IN ITERATION 1

1. Iteration	Enriched with common language media files...			
BPMNEasy1.2 modeling element	Doc	Image	Video	Audio
Entering request		x*1	x*2	
Enriching request info	x*3			
Informing requester			x*4	

The “Easy Capture Sheet” lists all BPMN^{Easy1.2} elements (table I, column 1). For instance, the first activity “Entering request” describes how a stakeholder has to enter the required data for a car parking slot request. Instead of using textual business rules or complex model constructs, the BPMN^{Easy1.2} element is enriched with common language media files. The attached image (table I, X*1) and the connected video sequence (table I, X*2) are used to explain the activity in detail. In addition to the description of the specific activities, the executing roles are added directly as a property to each activity. Furthermore, the color of the activities presents directly which kind of activity the stakeholders have to execute: manual (green form), semi-automated (blue form) and automated (red form). After finishing the “Easy Capture Sheet”, the modeled business process can already be executed (e.g. in a test run) by the responsible stakeholders. As stated in the agile methodology, following each iteration, a useable “product” must be available (although it is not completely finished). Before and after the execution, feedback was collected from the stakeholders. One feedback was that after entering the data, there are eight alternatives possible depending on the branch, location and type of the parking space request. The feedback is documented in the “Easy Capture Sheet” again. Table 2 shows an extract of the result of the second iteration.

TABLE II. EASY CAPTURE SHEET IN ITERATION 2

2. Iteration	Enriched with common language media files...				Feed-back
	Doc	Image	Video	Audio	
BPMNEasy 1.2 modeling element					
		X*1	X*2		
				X*3	More alternatives possible
				X*4	
			X*5		
...					

Distinguishing from the first iteration, a new element in form of an exclusive gateway (XOR) has been added (table II, row 3). A XOR defines that within a business process instance only one path can be taken [35]. Instead of

modeling all possible alternative business processes paths graphically, the alternatives have been captured in common language. After this modification, the business process was executed again.

The required information and data have been captured according to an agile method, e.g., to prove the correctness, stakeholders interacted and communicated with each other closely. In the following section, the usage of common language with an agile environment will be described in detail.

V. ENHANCED STAKEHOLDER SOCIALIZATION USING COMMON LANGUAGE IN AGILE BPM

According to Schienmann [37], three language types can be distinguished as follows:

- (1) The language of stakeholders (common language) as a problem-oriented language to communicate the requirements to an application system/business process, which has to be developed.
- (2) The language of BPM experts, e.g., diagrams languages such as BPMN 2.0, which are solution-oriented to model the requirements of the stakeholders with respect to the IT experts.
- (3) The language of IT experts, e.g., programming languages such as Java in order to realize, e.g., an automated business process.

This paper focuses in particular on human orientation and the usage of common language. For instance, according to [38] the quality of the (graphical) representation of business process models depends on the modeling experience of the stakeholders. Differences lead to errors and misunderstandings if there is no enhanced way of communication and involvement. Corresponding to Ortner's medial-real world model [39], this problem can be solved by improving the connection between the medial and real world. Within the medial world, requirements are defined which will be executed in the real world and monitored by the medial world again. The underlying architecture of Computational Society Architecture, in which stakeholders and IT systems are combined, has been depicted in Fig. 1.

The following sections describe how common language can be used to improve the collaboration and involvement of all stakeholders in agile environments.

A. Agile BPM

Various methods introduce agile BPM, e.g., [4][5]. In this section, the agile method BPM(N)^{Easy1.2} is used to present how and when during the methodology common language can be very helpfully applied. BPM(N)^{Easy1.2} consists of a modeling language (BPMN^{Easy1.2}), an approach (which explains, e.g., the interaction between different stakeholders) and of a tool called BPM Touch. BPM(N)^{Easy1.2} follows the method term definition of Ortner [41].

This combination allows the usage in all steps of Business Process Management – from business process capturing and modeling through analyzing, automation/execution and monitoring without losing its focus on integrating all stakeholders. Furthermore, it is possible to cover all layers of the Computational Society Architecture (Fig. 1).

- Language BPMN^{Easy1.2}

The elements set (graphic elements which can be used for modeling) of BPMN^{Easy1.2} specifies the BPMN 2.0 element set to a compact number of intuitive elements. BPMN^{Easy1.2} admits only elements which are generally known in common language – simple events (start, end, intermediate), simple gateways (AND, OR) and tasks/activities. For instance, there is an AND symbol which corresponds to the common understanding of “and”. In addition to each graphical element, media files can be added. For example, it is possible to add a video sequence to an activity to describe it in more detail. The compression at the level of modeling does not affect the XML Schema Definition (XSD) of BPMN 2.0. Every BPMN^{Easy1.2} model is stored in the form of BPMN 2.0. The BPMN 2.0 Data Object is used to keep a record of possible media files [4].

- Approach

Two connected cycles build the path of interaction and synchronization between all stakeholders. One cycle is used to capture or enhance business process models such as BPMN^{Easy1.2}. At the beginning of an iteration, BPMN^{Easy1.2} models are created. All stakeholders define the sequence of the captured activities, gateways and events. The BPMN^{Easy1.2} models are especially used to design the flow in general. If necessary, e.g., to describe an activity more in detail or to store a complex business rule, media files are used to add the additional information. All information is recorded in common language which makes it understandable for everybody. The modeled and formulated requirements can be the basis for modeling and implementation of enriched BPMN 2.0 business processes. This enrichment can lead to an automated workflow application. Therefore, the responsible stakeholders select a BPMN^{Easy1.2} model and a number of elements to work on. Parallel to the more technical work, stakeholders can directly start to create documentations or trainings. Within the flow path (within the iteration) quality gates are included. These quality gates are used to ensure that all stakeholders approve that the result, e.g., an automated business process, corresponds with the BPMN^{Easy1.2} models and captured media files (synchronization and acceptance). In addition, the second cycle describes the steps of analyzing/execution and optimizing. Predefined key performance indicators can be used for evaluating the stakeholders’ feedback, which have been collected within the business process execution. Immediately after the acceptance, new elements are captured or selected.

- BPM Touch tool

The mobile application BPM Touch follows innovative usability concepts. The focus is on user friendly features and the usage of mobile potentials. The modeling and user interface supports a revolutionary option for modeling business processes on mobile devices. For instance, after a business process has been selected by a simple touch on the sidebar, the business process model appears and can be directly edited. The flexible navigation is completed by a menu bar on the top, which provides basic functions to, for instance, create or save a new business process and by a dynamic pie menu to model a process flow very rapidly. Furthermore, media files can be assigned to every element of BPMN^{Easy1.2}. Audio files, which record, e.g., an oral description of an activity, video sequences, images and files, can be attached. Therefore, BPM Touch automatically loads the appropriate device, e.g., digital camera for video sequences. In addition, a “share”-button allows the direct distribution of BPMN^{Easy1.2} models to all stakeholders. For instance, the models can be exported to Microsoft PowerPoint for summarizing the complete documentation into a PowerPoint presentation. Summarizing the BPM(N)^{Easy1.2} method is based on the following foundations (c.f. [16]) :

- Common language, modeling languages, and programming languages

The common language is used throughout the entire business process management to communicate with all involved stakeholders. For instance, it can be necessary to transfer a business process model into a programming language, for example, to implement an automated workflow application. This is done by using agile concepts and focusing on interaction and synchronization in common language.

- Medial and Real World

Due to the iterative and incremental approach, in the context of BPM(N)^{Easy1.2} the medial/digital control level is closely linked to the level of doing (real world). Hereby, e.g., misinterpretations of stakeholders can be identified and corrected quickly.

- Terms and anchor

Dealing with term defects, such as synonyms, homonyms or false identifiers is intuitively supported by the usage of common language modeling. For instance, a transfer to other modeling languages is considerably simplified due to the iterative increased understanding of all involved stakeholders. Three anchors are taken as the primary goals of common language modeling: interaction, synchronization and quality. For example, the anchor of synchronization builds the organizational basis for a structured and proper coordination of all stakeholders involved.

However, the higher interaction and synchronization within agile methods can only lead to an enhanced stakeholder involvement if the project content is to be understandable for everybody. Otherwise, it may lead to quality issues of business processes [42]. Indeed, for some reasons, there is not enough time or effort spent for the required stakeholders coordination and communication. Counteracting this gap, e.g., to guarantee the quality of business processes, the following section explains different options to enhance the involvement of stakeholders during the application of an agile method to reach an enhanced implementation of the Computational Society Architecture.

B. Socialization of stakeholders

Fig. 2 illustrates the BPM(N)^{Easy1.2} approach and marks the parts in which stakeholder involvement is specifically enhanced:

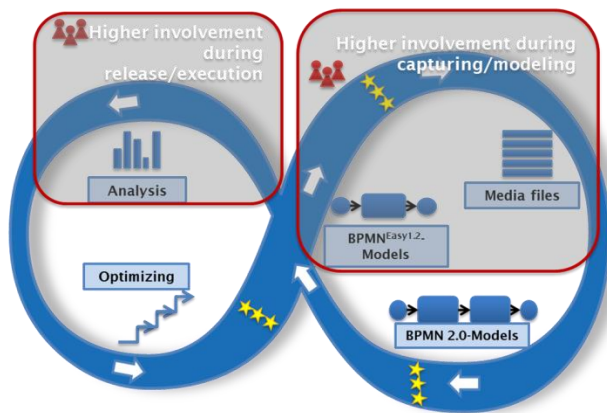


Figure 2. Illustration of the BPM(N)Easy1.2 approach

On one hand, stakeholders can be directly involved during the capturing and modeling of the business processes. On the other hand, BPM(N)^{Easy1.2} focuses on the increased involvement while executing or analyzing the documented business processes. The areas of optimizing and enriching the business processes models, e.g., towards BPMN 2.0, are omitted because of required expert knowledge which cannot be expected from all stakeholders. For instance, to automate business processes a higher technical skill is needed.

The approach described in Fig. 2 has been applied to the scenario presented in Section IV. Several aspects of an enhanced stakeholder involvement were identified:

- Enhanced documentation and publication

Within BPM(N)^{Easy1.2}, common language is used to simplify the documentation and publication of business process models. In fact, there are stakeholders who are not able to understand a graphic modeling language completely. To counteract this, the business process models can be exported or published in views which are comprehensible intuitively. For instance, all required information can be displayed as shown in table 1 and table 2. In addition, the BPM Touch application provides a clear overview of the

business process model elements. The following screenshot presents the user interface, which displays all information on the business process model element.

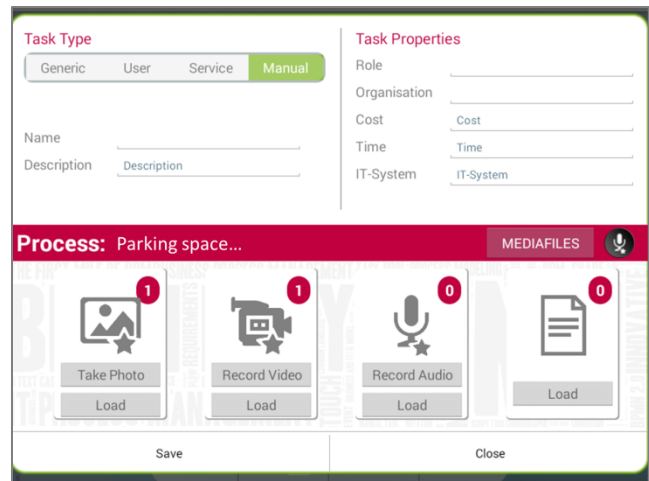


Figure 3. Screenshot of BPM Touch media file overview

As shown in Fig. 3, all collected common language media files are displayed at the bottom of the screen and can be opened by one touch easily.

- Enhanced capturing of information

Concerning capturing of information, BPM(N)^{Easy1.2} offers innovative ways. For instance, instead of modeling all paths of an OR gateway, BPMN^{Easy1.2} uses audio or video sequences, in which common language is used to explain the required behavior of the executing stakeholder. Fig.4 illustrates the difference between a traditional notation and BPM(N)^{Easy1.2}.

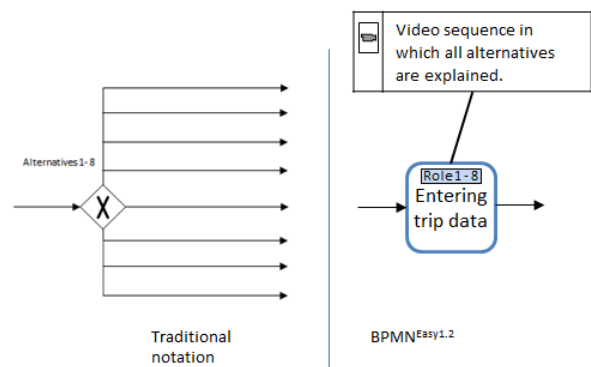


Figure 4. Illustration of modeling an OR gateway

As presented in Fig. 4, BPM(N)^{Easy1.2} models are highly compressed. Compared to traditional notations, in which already few alternatives can be confusing, the flow is still clear. In addition, the active recording of stakeholders leads to a higher involvement within the modeling phase.

- Enhanced quality

Gebhart et al. describe how quality of business processes can be improved and assured [42]. According to

this, the scenario described two iterations. Already in the second iteration, the collected feedback was taken to optimize the business process model. The usage of common language requires no training and motivates stakeholders to integrate their feedback. Predefined quality attributes are taken to check the expected quality. Furthermore, the quality attributes can be formulated in common language.

- Enhanced interaction using BPM(N)^{Easy1.2} role concept

The usage of concrete BPM roles improves the interaction and synchronization within a BPM project. The screenshot in Fig. 5 presents a view of modeling a business process model on BPM Touch.

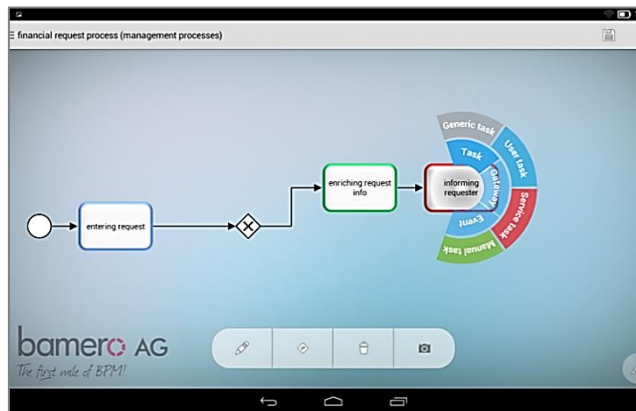


Fig. 5. Screenshot of BPM Touch modeling perspective

BPM(N)^{Easy1.2} adapts (c.f. Scrum [14]) and introduces three roles: BPM expert, IT expert and (key) user. In Fig. 5, the BPM expert uses the BPMN^{Easy1.2} model to capture all information and synchronize all roles in predefined time slots. Furthermore, the BPM expert instructs the other stakeholders. Hereby, IT experts are in charge of required software implementations and (key) users coach each other or execute the business processes. For communication, all roles use common language so that everybody is able to be involved instead of excluded. Potential misunderstandings, e.g., different interpretations of a specific term will be caught and solved in the next iteration. The BPM expert accompanies and promotes the discussion between all stakeholders. With the help iterative adjustments, e.g., the business process models or the user interfaces of an application, the optimization will be done.

VI. CONCLUSION AND OUTLOOK

In this paper, common language has been applied in an agile BPM environment to promote the involvement of all stakeholders by increasing the general access and understanding of modeled business processes. For this purpose, we exemplarily chose the agile BPM method BPM(N)^{Easy1.2} and applied it to a concrete scenario. BPM(N)^{Easy1.2} focuses on describing a method which delivers an approach, a language and a tool for business process management in an agile environment. The existing methods are enhanced by using common language for

interaction, synchronization and quality assurance based on the perspective of the Computational Society Architecture.

The scenario showed how to use techniques such as the “Easy Capture Sheet” and how to apply the innovative BPM Touch mobile application. BPM Touch can be used for capturing all required information for business process models. Furthermore, it has been shown that BPM Touch is also useful to increase stakeholder involvement, e.g., in the execution and publication phases of business processes. All these aspects lead to “living” business processes instead of creating rigid documentations. However, the investigation of more aspects using common language business process management has to be a subject of further research. Moreover, more projects such as the modeling project of “Smart City Constance” have to be initialized to validate the method and concept of common language and social stakeholder involvement.

REFERENCES

- [1] T. M. Bekele and Z. Weihua, “Towards collaborative business process management development current and future approaches”, in Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference, 2011, pp.458-462.
- [2] W. M. P. van der Aalst, A. H. M. ter Hofstede and M. Weske, “Business Process Management: A Survey”, In: van der Aalst, W. M. P.; ter Hofstede, A. H. M.; Weske, M. (Hrsg.): Proc. Intl. Conf. on Business Process Management (BPM 2003), Lecture Notes in Computer Science, Vol. 2678, S. 1–12. Springer Verlag, Berlin, 2003.
- [3] M. Weske, 2007, “Business Process Management: Concepts, Languages, Architectures”, Potsdam : Springer Verlag Berlin Heidelberg, 2007.
- [4] M. Mevius and P. Wiedmann, „BPM(N)Easy1.2 – Gebrauchssprachliche Gestaltung IT-basierter Prozesse. BSOA 2013. 8. Workshop „Bewertungsaspekte service- und cloudbasierter Achitekturen“ der GI Fachgruppe „Software-Messung und -Bewertung“, Basel, 2013, pp. 31-46.
- [5] F. Schnabel, Y. Gorrongoitia, M. Radzimski, F. Lecue, N. Mehandjiev, G. Ripa, et al. “Empowering Business Users to Model and Execute Business Processes”, BPM 2010 Workshops, Springer, 2011, pp. 433-448.
- [6] Fraunhofer IESE, “Marktanalyse zu BPM Suites”, http://www.iese.fraunhofer.de/content/dam/iese/de/dokumente/oeffentliche_studien/Fraunhofer_IESE_Studie_BPM-Suites2013.pdf, accessed on 05/21/2014.
- [7] A. Fleischmann and W. Schmidt, “Business Process Monitoring with S-BPM”, S-BPM ONE - Running Processes Communications in Computer and Information Science Vol. 360, 2013, pp 274-291.
- [8] B. Weber and W. Wild, “An Agile Approach to Workflow Management”, Lecture Notes in Informatics, Modellierung 2004, ISBN 3-88579-374-1, pp. 187-202.
- [9] K. Bergener, J. vom Brocke, S. Hofmann, A. Stein and C. vom Brocke, „On the importance of agile communication skills in BPM education: Design principles for international seminars”, Knowledge Management & E-Learning: An International Journal (KM&EL), ISSN 2073-7904, EISSN 2073-7904, 2012.
- [10] S. Nurcan and R. Schmidt, “BPM and Social Software”, BPM 2008 Workshops, LNBIP 17, 2009, pp. 625–634.
- [11] S. Bittmann and O. Thomas, “A theory of practice modelling – Elicitation of model pragmatics in dependence to human actions”, Lecture Notes in Informatics, Modellierung 2014, ISBN 978-388579-619-0, p.33-48.
- [12] J. Highsmith, “Agile Software Development Ecosystems”, Pearson Education, 2002, pp. 26-34.
- [13] Agile Manifesto, <http://agilemanifesto.org>, accessed on 05/21/2014.

- [14] K. Schwaber and J. Sutherland, "The Scrum Guide, The Definitive Guide to Scrum: The Rules of the Game", 2011, http://www.scrum.org/Portals/0/Documents/Scrum%20Guides/Scrum_Guide.pdf, accessed on 05/21/2014.
- [15] K. Beck, "Extreme Programming Explained: Embrace Change", Addison-Wesley Longman, 1999.
- [16] W.M.P. van der Aalst, "Business Process Management: A Comprehensive Survey", Hindawi Publishing Corporation ISRN Software Engineering Vol. 2013, <http://dx.doi.org/10.1155/2013/507984>.
- [17] M. Mevius, E. Ortner and P. Wiedmann, "Gebrauchssprachliche Modellierung als Grundlage für agiles Geschäftsprozessmanagement", Lecture Notes in Informatics, Modellierung 2014, ISBN 978-388579-619-0, p.169-184.
- [18] A. Komus, „BPM Best Practice Die wichtigsten Erkenntnisse aus aktuellen Praxis Studien Auf dem Weg zu einem ganzheitlichen BPM 2010, <http://www.komus.de/fileadmin/downloads/public/2010-DSAG-BPM.pdf>, accessed on 05/21/2014.
- [19] R. C. Mayr, "Keynote to Modellierung 2014", Vienna, 2014.
- [20] B. Hauser, "Optimierung von Prozessen in der internen Kommunikation", Speech on 9th Process Solution Day, Cologne, 2014.
- [21] E. Heinemann, „ Sprachlogische Aspekte rekonstruierten Denkens, Redens und Handelns – Aufbau einer Wissenschaftstheorie der Wirtschaftsinformatik“, Gabler Verlag, Wiesbaden, 2006.
- [22] M. Parameswaran and A. Whinston, „Social Computing: an overview“, Communications of the Association for Information Systems. 2007, Vol. 19, pp. 762-780.
- [23] H. Wedekind and E. Ortner, „Systematisches Konstruieren Von Datenbankanwendungen: Zur Methodologie Der Angewandten Informatik“, Herrn Prof. Dr. K. F. Bußmann Zum 65. Geburtstag. München [u.a.] : Hanser, 1980.
- [24] B. Eller, „Usability Engineering in Der Anwendungsentwicklung: Systematische Integration Zur Unterstützung Einer Nutzerorientierten Entwicklungsarbeit“, Springer, 2009.
- [25] E. Ortner and B. Schienmann, Normative language approach a framework for understanding. Conceptual Modeling — ER '96, Lecture Notes in Computer Science Vol. 1157, 1996, pp 261-276.
- [26] D. L. Moody, „The “Physics” of Notations: Toward a Scientific Basis for Constructing Visual Notations in Software Engineering“, Ieee transactions on Software Engineering, vol. 35, no. 6, November/December 2009, p.756-779.
- [27] E. Evans, "Domain-driven design: Tackling complexity in the heart of software", Addison-Wesley. Boston and MA [etc.], 2003.
- [28] M.Cohn, "UserStories", <http://daviding.wordpress.com/2013/07/27/mike-cohn-user-stories/>, 2012, accessed on 05/21/2014.
- [29] M. Glinz, "Very Lightweight Requirements Modeling, " presented at the Requirements Engineering Conference (RE), 2010 18th IEEE International, 2010, pp. 385-386.
- [30] H. J. Schmelzer and W. Sesselmann, „Geschäftsprozessmanagement in der Praxis“, Carl Hanser Verlag, München, 2008.
- [31] F. Schnabel, Y. Gorrongoitia, M. Radzinski, F. Lecue, N. Mehandjiev, and G. Ripa, "Empowering Business Users to Model and Execute Business Processes," BPM 2010 Workshops, Springer, 2011, pp. 433-448.
- [32] P. Antunes, D. Simões, L. Carriço and J. A. Pino, "An end-user approach to business process modeling", Journal of Network and Computer Applications 36 (2013), p. 1466 -1479, 2013.
- [33] S. Bittmann, D. Metzger, M. Fellmann and O. Thomas, „Additional Information in Business Processes: A Pattern-Based Integration of Natural Language Artifacts“, Lecture Notes in Informatics, Modellierung 2014, ISBN 978-388579-619-0, p.137-152.
- [34] G. Bruno et al., "Key challenges for enabling agile BPM with social software", in Journal of Software Maintenance and Evolution: Research and Practice, 2011, 23; S. 297–326.
- [35] OMG, "Business process model and notation (BPMN)", Version 2.0, 2011.
- [36] Axon IVY AG, "BPM Suite Axon.ivy", <http://axonivy.com>, accessed on 12/03/2014.
- [37] R. Barker, "Case Method: Entity Relationship Modelling" Addison-Wesley Professional, 1990.
- [38] bamero AG, "BPM Touch® - The Revolution", <http://www.bamero.de>, accessed on 05/21/2014.
- [39] B. Schienmann, „Objektorientierter Fachentwurf: ein terminologiebasierter Ansatz für die Konstruktion von Anwendungssystemen“, B.G. Teubner Verlag, Stuttgart, 1997.
- [40] L. Burmester, Adaptive Business-Intelligence-Systeme, DOI 10.1007/978-3-8348-8118-2_3, Vieweg+Teubner Verlag Springer Fachmedien Wiesbaden GmbH, 2011.
- [41] E. Ortner, „Semantisch normierte Anwendungssysteme und die >>eingeschränkte Freiheit<< der IT-Nutzer“, in Mittelstraß, J.: Zur Philosophie Paul Lorenzens. mentis Verlag, Münster, 2012.
- [42] M. Gebhart, M. Mevius and P. Wiedmann, „ Application of Business Process Quality Models in Agile Business Process Management“, The Sixth International Conference on Information, Process, and Knowledge Management, ISBN: 978-1-61208-329-2.

A Versioning and Commenting Approach for Enhancing Group Efficiency in Collaborative Web-Based Business Process Modeling Tools

Justus Holler

University of Muenster

ERCIS

Muenster, Germany

justus.holler@ercis.uni-muenster.de

Abstract—Business process modeling interpreted as a collaborative act requires resource intense communication and coordination between domain and modeling experts. Therefore, modern web-based business process modeling tools need to provide a shared workspace. Tool users can check and validate the progress of the business process modeling project and coordinate their work. This paper proposes a concept for workspace awareness, which combines version management with a commenting functionality in order to increase the groups modeling efficiency. The process model versions in the workspace from draft to final are documented with the help of a history-in-parent approach. Comments for each model version and comments within the models allow the tool users to collaborate asynchronously, which decreases the need for several workshop and interview iterations for process recording and refinement.

Keywords—Business Process Modeling Tool; Versioning; Comments; Awareness; Modeling Efficiency.

I. INTRODUCTION AND MOTIVATION

Business Process Management (BPM) deals with managing, transforming and improving organizational operations [1]. One key part is the modeling of these operations in business process models. They are of high interest for public and private organizations as means to stay competitive and attractive in fast changing markets. A variety of methods, modeling languages and tools are used in these endeavors and there is research regarding the suitability of modeling languages for different purposes [2][3]. In every case, a modeling project basically depends on two parties. Firstly, the department worker who holds the knowledge of the organizational domain. Secondly, the modeling expert with particular sharp analytical and modeling skills who facilitates model creation [4-9].

Projects in this field are typically resource intense. This applies on the one hand to the time the department workers have to invest into the workshops or interviews and on the other hand to the budget for the (external) modeling expert who moderates the workshops or interviews and designs and refines the models afterwards [10][11]. Subsequently, the designed models have to be semantically validated by the department workers [6][7] and finalized in an iterative manner.

In order to reduce errors and to enhance the holistic understanding of the business processes within the scope of the project, it is at least beneficial if not necessary to include several department workers or domain experts into the business process modeling project [12][13]. Involving department workers in the process of modeling is reasonable as they accept [12][14] and understand the processes more thoroughly which fosters their critical evaluation and leads to potential process improvements [15]. Furthermore, the risk is reduced that the designed processes are not accepted by the department, which would lead to resource intense re-modeling or a failed project.

In its core, a business process management project with the deliverable of a documented process landscape is based on communication. Project participants contribute to a project outcome no single one of them could solely have accomplished. The necessary integration of the relevant project members can be done with the help of collaborative business process modeling tools [9][14]. In such tools, collaboration is possible but also coordination is necessary as the project participants have to distribute their tasks and synchronize their working objects [16-18] in a common context. Within this context awareness information helps to inform about the task status of other project members tasks and therefore reduces the coordination effort [19][20].

Hence, it is worthwhile to identify and foster drivers for improved integration and communication of the project members. The overall goal of this paper is to increase the project groups' efficiency in designing, reviewing and finalizing the models in the business process landscape. Therefore, this paper proposes a concept for workspace awareness in web-based business process modeling tools. This is done by the combination of a model versioning approach and an integrated commenting function.

Section 2 examines the need for versions in business process modeling projects and the role of comments as means for communication and coordination in information systems. The third Section explains the reasonable combination of model versioning and comments for improving awareness and presents the history-in-parent versioning approach and how it can be applied for a web-based process modeling tool. Section 4 discusses the conceptual design of the management and visualization of the commenting function based on the versioning approach

and Section 5 concludes the paper and gives an outlook for future research.

II. VERSIONS AND COMMENTS IN BUSINESS PROCESS MODELING TOOLS

In the context of modeling, it is difficult to define the quality of process models. Models can only be more or less useful with respect to their purpose [21]. In order to reach a status of usefulness – a final model – it needs the discussion of the good-willing and knowledgeable [22]. Moreover, once this status is reached, it has to be taken into account that organizations, markets and requirements change and that the process models have to be adjusted or re-evaluated on a continuous basis. In professional life in process modeling projects, this is done with IT support. For example, with a web-based process modeling tool which supports the project members in creation, modification, validation and communication of the process models in a collaborative manner. The project members access the models in an online repository in parallel and will either only view the model or also modify their content. The synchronous modification bares the risk of known issues like lost updates or inconsistencies within the models if not appropriate techniques are applied to reduce or prevent the risks. In theory this risk can be managed by displaying the models in real-time but it is questionable if this can be achieved and guaranteed in practical settings with tool users working distributed regarding time and place. Hence, (theoretical) modeling tools with the support for synchronous modeling are not applied in practice [23] and an analysis regarding twelve commercial available modeling tools showed that none of them supported real-time modeling [10].

Modeling in professional life is done in an asynchronous manner. The modeling experts design versions of the model, while the domain experts contribute business knowledge and validate the model versions either in workshops or directly within the tool. In team settings with several domain and modeling experts, there will be discussion regarding the most reasonable model and hence, the need for different versions of one process model. It takes several distinctive versions to reach the final, correct and useful state of one model.

For each new version from draft to the final model of the whole process landscape the reason why the new version was created, e.g., the input by the domain expert has to be saved. Hence, the necessity for textual descriptions or additional documents arises. This meta-information helps to understand what has changed or why something should be changed in a new version of the model. Therefore, comments enhance the understanding of the model users collaborating in one (virtual) workspace regarding the reasons for model changes and reduce the coordination effort [19].

III. WORKSPACE AWARENESS WITH COMMENTS AND VERSION MANAGEMENT

People who work together in one room can easily recognize who is dealing with which part of the project. In a distributed or virtual setting, this is not possible. It needs additional information for efficient coordination of work

items. Awareness information addresses the “what” and “where” [19]. In the case of comments, the comment itself is the answer to the “what” while the link to the commented object answers the “where”. Comments within process models therefore are particular well-suited regarding awareness, as their information automatically sets a context and the context for the process model element does not have to be set by a user explicitly. In groupware, awareness is a long known research field [24] and in the area of software engineering comments are used because of their awareness effects and are the basis for code awareness or code repository mining [25–27]. Although it is expected that the benefits of the above mentioned collaborative tasks can be transferred to the other collaborative tasks [20][28], there is no particular research in the area of collaborative process modeling with a focus on comments. Only [29] considers the commenting function as possible interface between domain and modeling experts. They can use comments directly in the process model (review comments) as known from software like Adobe Reader or Microsoft Word which allow a precise validation and correction and a dialogue between the users [20]. Domain experts can add information, which was not recorded during the interview or workshop. If the modeling expert creates new versions based on the input and wants to communicate the background for the new version, a management system has to be in place allowing comments (version comments) as meta-information for the specific versions.

A. Version Management Systems

The main purpose of version management systems in software engineering is the management of different versions of textual documents (source code) and their consolidation [30]. The most important version management systems like Concurrent Versions System (CVS), Subversion (SVN) and Git incorporate textual descriptions for the specific version or code which is a feature regularly used by software developers [27][31]. The developer does not have to analyze the source code in detail and still can estimate the impact of the changes for his or her work. The same applies to users of web-based business process modeling tools. Domain and modeling experts are aware of the changes made to one model version (version comment) or to a specific model element (review comment) by reading the textual description.

A hurdle to overcome for the application of version management concepts in web-based business process modeling tools is the document format in which process models are stored. Unlike source code in software engineering, process models cannot be compared on the basis of lines of text. Business process models are typically not stored in text files but in a binary format or in the best case in a serialized XML format [32]. Hence, process modeling tools which have an integrated version management functionality usually depend on their proprietary format [10].

The “History-in-Parent” version management approach [33] is proposed here which can directly be applied on a database level. Thus, there is no need for a serialization of

the process models. The pre-condition for this approach is that the business process landscape can be represented as tree structure with the process framework as root node and the main and detail processes with their respective process elements as child nodes [34]. In principle, the level of abstraction layers is arbitrary but for readability reasons this paper limits the levels of abstraction to three. As only the history (old versions) is stored, the displayed process is always the most recent version of the model, which leads to good processing time in accessing the model in a web-based context. Also the processing time for executing basic functions as creating, modifying or deleting process elements (nodes) or restoring an old version is nearly optimal [33].

In Figure 1, an example structure is shown with a process framework (root node of the tree structure and 1st abstraction layer) after four modeling steps ($t = 4$) with links to a “1st main process” and “2nd main process” (both child nodes of the process framework). The “1st main process” is further detailed by a “1st detail process” and “2nd detail process”. Hence, the “1st main process” is the parent node for the two detail processes. For each layer of parent nodes (two in this example), a table is maintained, which stores the historic connections of the parent node to its child nodes. The “historic tables” are represented in Figure 1 as smaller circles in the same color and line style as the respective level of abstraction. The naming of the history nodes in the figures follows on every layer the same principle: The first part of the identifier represents the ID of the parent node (in the example the “Process framework” or “1st main process” with the ID 1). The second part of the identifier after the dot represents the time when the connections form parent to child nodes were valid.

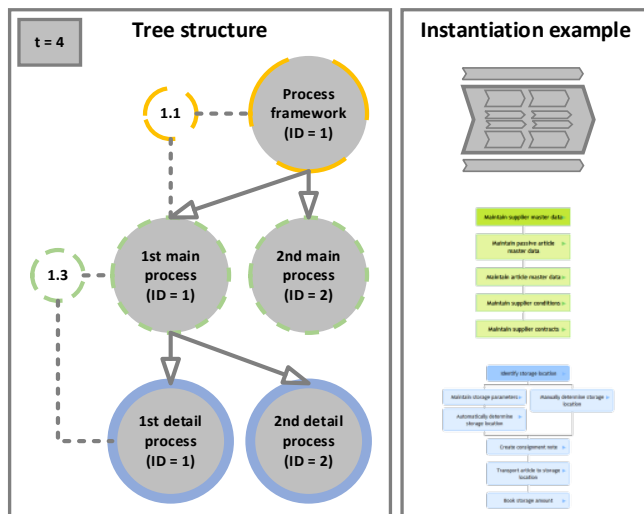


Figure 1. Tree structure and instantiation example of nodes.

Table 1 and Table 2 show the relevant data base tables for the main processes layer of the example. In Table 2, the historic information of the yellow history node can be found in the last row. For readability, the history table for the detail processes which would be analog to Table 2 is not

printed and the visualization of technical attributes like primary and foreign keys are omitted in these and all following tables. Also, non-relevant history nodes are not displayed in the figures.

TABLE I. MAIN PROCESSES DATA BASE TABLE

Name	Node-ID	Parent-ID
1st main process	1	1
2nd main process	2	1

TABLE II. HISTORY TABLE ON THE PROCESS FRAMEWORK LEVEL

Framework-ID	Child nodes	Validity
1	NULL	0
1	1	1

B. Creating, Modifying, Deleting and Restoring Process Elements

A version management approach in a modeling tool has to track all relevant changes made to a model by the basic operations create, modify or delete. Furthermore, restoring old model versions has to be supported. In the following, the necessary steps for each of these operations are explained and an example is given in the next subsection.

Saving the links between parent node and its child nodes in the respective history table is the first step for all operations. Three attributes are of relevance: the ID from the parent node k , the IDs of the *child nodes*, which were linked to k and the version number (*validity*) before the change took place.

In case a node o is *modified*, its parent node is saved as described in the previous paragraph. Then, node o is copied as new node k with the same parent ID but the modified values (e.g., new identifier). The ID of the new node k is (auto-)incremented by the database system. Now, the parent node ID of o is set to NULL because o has no valid connection to a parent node in the most recent version of the model. Then o is saved and all links to o from o 's child nodes have to be changed to links to k .

If a new node k is *created*, first the parent node of k is saved so the old links are saved for the validity $t - 1$ and then the new node k can be created and is linked to the parent as new process element in the most recent model version.

The procedure of *deleting* a node k is done with the following steps: at first the parent node of k is saved, than the attribute parent node from k is set to NULL and k is saved. Finally, for all child nodes the links to k are set to NULL because k does not exist anymore in the most recent version of the process landscape.

Restoring a model version x is always done relatively to a node in the process landscape. The node (v) which shall be set back to version x is treated as root of a sub tree. This tree is then restored with the connections valid for the specified version (x): first, the parent node of v is saved (if existing) and afterwards v itself is saved. In a recursive step the original o of v valid for the time x has to be identified. Therefore, in the history table of v the most recent node h has

to be selected where v is not listed as child node. Three cases are possible: If no node h is found, then, there was no change to v since version x . If h is found and the next historic child node $h+1$ in the table has the same amount of child nodes as h , then the original o can be derived by the comparison of the child nodes from h and $h+1$. In case $h+1$ has only one more child node, than v is not the result of a modification but the original itself and the recursion can stop. After going through the recursion at least twice the value for the parent node of v is copied to o and the parent node of v is set NULL, which makes v to the new original. All child nodes of v have to be saved and their parent node attribute is set to NULL. Now, search for the newest historic node h of v in the historic table of the same level like v which version is smaller or identical to x and set the value for the parent node of all in h listed child nodes to v . For clarification of the above mentioned steps, all operations are used in the following modeling scenario.

C. Modeling an Example Process Landscape

In the first step ($t = 0$), a new process framework “Whole sale” is created (Figure 2).



Figure 2. Process landscape ($t = 0$).

After creating the main processes contracting ($t = 1$), purchasing ($t = 2$) and receiving ($t = 3$) and the detail processes maintain article master data ($t = 4$), maintain supplier master data ($t = 5$) and maintain supplier contracts ($t = 6$) the process landscape in its tree representation looks like depicted in Figure 3 with the process element nodes and the respective history nodes.

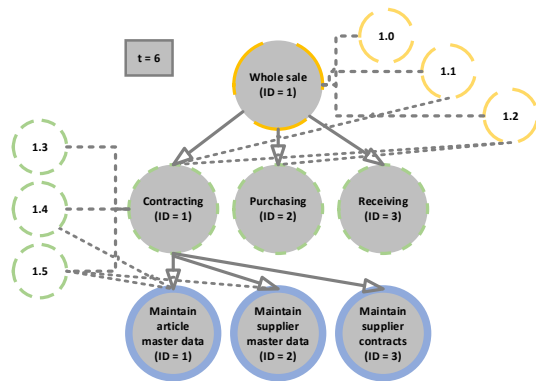


Figure 3. Process landscape after adding main and detail processes ($t = 6$).

Now the main process element “Contracting” is renamed to “Contrac management”. The renamed element is linked to the parent node while the node with the old name (“Contracting”) has no link to a parent node anymore (Figure 4, $t = 7$).

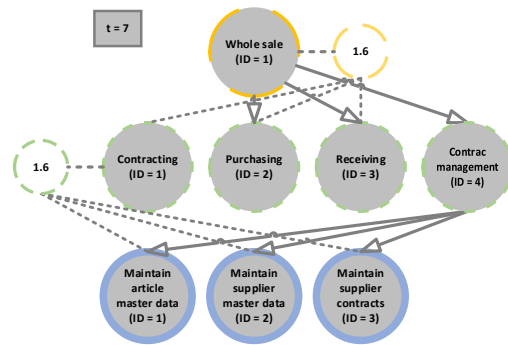


Figure 4. Process landscape after renaming Contracting ($t = 7$).

TABLE III. MAIN PROCESSES AFTER CHANGING “CONTRACTING” TO “CONTRAC MANAGEMENT” ($T = 7$)

Name	ID	Framework-ID
Contracting	1	NULL
Purchasing	2	1
Receiving	3	1
Contrac management	4	1

Due to the name change a new entry in the table of the main processes (Table 3) is added (ID = 4) with a new connection to the old parent of “Contracting”. The old connection, which was valid for $t = 6$ is stored in the history table (Table 4).

TABLE IV. HISTORY TABLE ON FRAMEWORK LEVEL AFTER CHANGING “CONTRACTING” ($T = 7$)

ID	Child nodes	Validity
1	NULL	0
1	1	1
1	1,2	2
1	1,2,3	6

Table 5 lists all detail processes including the ID of their parent nodes. In $t = 7$, the parent node ID for the detail processes is the same as they all have the same parent node which has changed from ID 1 (“Contracting”) to ID 4 (“Contrac Management”).

TABLE V. DETAIL PROCESSES DATA BASE TABLE ($T = 7$)

Name	ID	Main proc.-ID
Maintain article master data	1	4
Maintain supplier master data	2	4
Maintain supplier contracts	3	4

The version information regarding the old parent ID is stored in the historic table on the main process level as this is the parent-level of the detail processes (Table 6).

TABLE VI. HISTORIC TABLE OF THE MAIN PROCESSES ($T = 7$)

Main proc.-ID	Child nodes	Validity
1	NULL	3
1	1	4
1	1, 2	5
1	1, 2, 3	6

A. Storing the Comments in the Data Base

All comments regarding the process landscape are stored in one comment table (Figure 7, [36]). The link to the specific model (element) is done via: a *comment-ID* identifies the specific comment, an *element-ID* sets the link to the process element or process which is commented (the process element table is omitted for readability reasons), an *author-ID*, an *addressee-ID* can be linked to a specific user, an *attachment* attribute allows a link to additional information, a *timestamp* allows to sort the comments, a *type* indicates if it is a version, review or change request comment and the attribute *comment* contains the textual comment itself. The hierarchical relationship of the comment entity (*comment hierarchy*) allows users to write comments in relation to another comment and therefore create answers or threads.

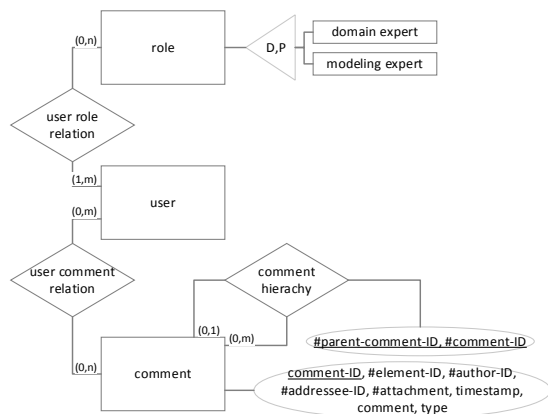


Figure 7. Entity relationship model of comment, user and role table.

Between the *comments* and the *users* a n:m-relationship is proposed. This allows a read / unread functionality for each user separately as known from e-mail clients.

The process and process element tables are not depicted in the Entity relationship model (ERM, [36]), but have an important status attribute. Each instance on framework, main and detail process level has a status indicating if the element is in work, a draft, ready for review or final. Depending on this status, the domain expert can validate the overall structure (draft), check for semantic correctness (ready for review) or set the status of the process (element) to final.

These status are coupled with the role concept. As in the modeling project domain and modeling experts are involved it has to ensure, that only authorized persons can set the appropriate status. For example, the modeler can set the “in work”, “draft” and “ready for review” status while only the domain expert can set the status to “final” and therefore validate its semantic correctness from the subjective point of view and if it is fit for use [21].

B. Communication of the Comments

In order to create workspace awareness for the project members it is crucial to communicate the committed changes, their textual descriptions (version and review comments) and the status of each comment. Hence, the

comments have to be *visualized*, and *managed* by the modeling tool.

In Figure 8 a domain expert views the main process contract management and has one unread comment regarding the process step “Maintain supplier master data”. The amount of unread comments is visualized with the help of a bubble. It can either relate to comments regarding the process element itself or it can indicate the amount of comments underneath the process element. If this bubble or the process element is right-clicked the user can navigate to the comments view (Figure 9), switch to the version management view (Figure 10) or change the process status. Furthermore, in Figure 8 the domain expert writes a change request regarding the missing process step “Cancel supplier contract”.

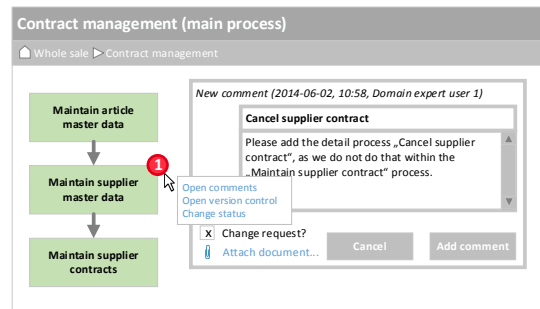


Figure 8. Adding a comment and right-click effect (domain expert role, t = 9).

While *viewing* the comments in the process model the user can mark the comment as read and depending on the comment and the users’ role, it is possible, to “accept” or “reject” a change request. By this means, a domain expert can track if the modeling expert already read the comment and if her change request was accepted and possibly send the modeling expert a reminder. Furthermore, the read / unread functionality helps to filter the visible information: all read and accepted or rejected comments are hidden in the default visualization of the process model for the specific tool user. If there is no new comment, there would be no red bubble in the process model visible. In Figure 9 the bubble indicates the two new comments. In the comment view the comments are displayed in a discussion like style and additional attached documents can easily be accessed by the user [28].

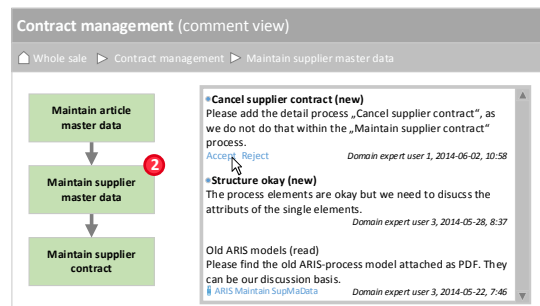


Figure 9. Comment view (modeling expert role, t = 9).

In Figure 10 the detail process “Cancel supplier contract” was added which leads to a new version with an explicit version comment by the modeling expert 2 user. Furthermore, the authorized user can restore old versions of the model from here.

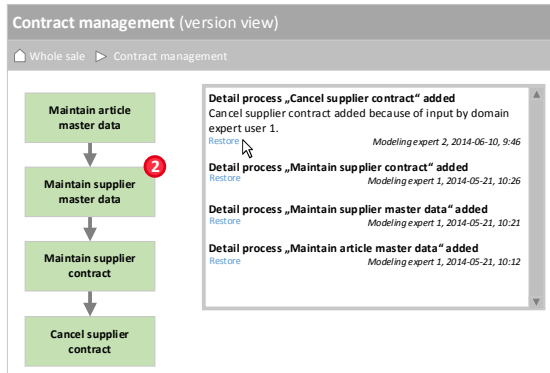


Figure 10. Version view (modeling expert role, t = 10).

Typically, the management of the comments is done in a passive manner and relies on the project member to log. After the user logs in, a dashboard indicates about activities in the modeling project with a list of recently changed models and new comments. However, for a domain expert, working fulltime in the department, this is problematic as their day to day business does not allow visiting the tool on a regular basis. Hence, for the efficient collaboration between the domain and modeling experts, the comments regarding changes have to be communicated also actively. This is achieved by sending requests for comments or requests for validation via e-mail. In addition, reports on a regular basis summing up all changes and indicating where new comments were made are a possibility of increasing the awareness.

V. CONCLUSION AND FUTURE WORK

This paper presented an approach for awareness enhancement in web-based business process modeling tools. The basis for this is a version management instantiation of the history-in-parent approach, which directly works on the database tables. The model versions can be commented and the modeling team is informed about the changes. As the goal of the approach is to foster the efficiency in the collaborative act of modeling the approach defines active (e-mail notification) and passive awareness (indicators for changes in the tool) functionality. Thus, the approach is striving for supporting domain and modeling experts likewise by reducing coordination effort, which leads to savings in time and therefore overall more efficient modeling.

In future research, this approach will be implemented in a web-based business process modeling tool for evaluation. The data base performance over time and feature completeness will be evaluated. Especially the medium (e.g., e-mail, dashboard), frequency (directly vs. summarized) and

mode (active vs. passive) of awareness information will be evaluated.

Also, the limitations, e.g., the technical limitation regarding the danger of inconsistencies in case of concurrent modeling will be addressed. One possibility of addressing this can be a locking mechanism. The evaluation will show to which extent this lock hinders the collaborative work, as restricted to the approach all models underneath the model in use would be locked for other modelers.

Furthermore, the evaluation in laboratory settings with students and afterwards in consultancy projects will reveal the appropriateness of the comments. As no user can be forced to actively read or write comments, there is a natural limitation in this approach. It is likely that convenience in writing the comments like templates or auto-completion, similar to the popular version management control systems, will foster comment quality.

REFERENCES

- [1] M. Hammer and J. Champy, “Reengineering the Corporation: a Manifesto for business Revolution,” *Bus. Horiz.*, vol. 36, no. 5, pp. 90–91, 1993.
- [2] P. Wohed, W. M. P. van der Aalst, M. Dumas, A. H. M. ter Hofstede, and N. Russell, “On the suitability of BPMN for business process modelling,” in *Business Process Management*, Springer, 2006, pp. 161–176.
- [3] R. S. Aguilar-Savén, “Business process modelling: Review and framework,” *Int. J. Prod. Econ.*, vol. 90, no. 2, pp. 129–149, Jul. 2004.
- [4] M. Pendergast, K. Aytes, and J. D. Lee, “Supporting the group creation of formal and informal graphics during business process modeling,” *Interact. Comput.*, vol. 11, no. 4, pp. 355–373, 1999.
- [5] D. Dean, R. Orwig, and D. Vogel, “Facilitation methods for collaborative modeling tools,” *Gr. Decis. Negot.*, vol. 9, no. 2, pp. 109–128, 2000.
- [6] P. J. M. Frederiks and T. P. Van der Weide, “Information modeling: the process and the required competencies of its participants,” *Data Knowl. Eng.*, vol. 58, no. 1, pp. 4–20, 2006.
- [7] M. Weske, *Business Process Management*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.
- [8] I. Wilmont, E. Barendsen, S. Hoppenbrouwers, and S. Hengeveld, “Abstract Reasoning in Collaborative Modeling,” 2012 45th Hawaii Int. Conf. Syst. Sci., pp. 170–179, Jan. 2012.
- [9] P. Rittgen, “Collaborative Modeling - A Design Science Approach,” in *HICSS '09. 42nd Hawaii International Conference*, 2009, pp. 1–10.
- [10] K. Riemer, J. Holler, and M. Indulska, “Collaborative Process Modelling - Tool Analysis and Design Implications,” in 19th European Conference on Information Systems (ECIS), Paper 39, 2011.
- [11] J. Jeston and J. Nelis, *Business process management*. Routledge, 2014.
- [12] D. Dean, R. Orwig, J. Lee, and D. Vogel, “Modeling with a group modeling tool: group support, model quality, and validation,” in *Twenty-Seventh Hawaii International Conference*, 1994, vol. 4, pp. 214–223.
- [13] T. Dollmann, C. Houy, P. Fettke, and P. Loos, “Collaborative Business Process Modeling with CoMoMod - A Toolkit for Model Integration in Distributed Cooperation Environments,” 2011 IEEE 20th Int. Work. Enabling Technol. Infrastruct. Collab. Enterp., pp. 217–222, Jun. 2011.
- [14] P. Rittgen, “Collaborative modeling of business processes: a comparative case study,” in *Proceedings of the 2009 ACM symposium on Applied Computing*, 2009, pp. 225–230.

- [15] F. M. Santoro, M. R. S. Borges, and J. A. Pino, "CEPE: cooperative editor for processes elicitation," in 33rd Hawaii International Conference on System Sciences, pp. 10 vol 1, 2000.
- [16] M. Stefik, G. Foster, D. G. Bobrow, K. Kahn, S. Lanning, and L. Suchman, "Beyond the Chalkboard: Computer Support for Collaboration and Problem Solving in Meetings," *Commun. ACM*, vol. 30, no. 1, pp. 32–47, 1987.
- [17] S. R. Fousseu, R. E. Kraut, F. J. Lerch, W. L. Scherlis, M. M. McNally, and J. J. Cadiz, "Coordination, Overload and Team Performance: Effects of Team Communication Strategies," in Proceedings of the 1998 ACM conference on Computer supported cooperative work, 1998, pp. 275–284.
- [18] C. A. Ellis, S. J. Gibbs, and G. L. Rein, "Groupware - Some Issues and Experiences," *Commun. ACM*, vol. 34, no. 1, pp. 39–58, 1991.
- [19] C. Gutwin, S. Greenberg, and M. Roseman, "Workspace awareness in real-time distributed groupware: Framework, widgets, and evaluation," *People Comput. XI*, pp. 281–298, 1996.
- [20] P. Dourish and V. Bellotti, "Awareness and coordination in shared workspaces," 1992, pp. 107–114.
- [21] J. Becker, M. Rosemann, and C. von Uthmann, "Guidelines of Business Process Modeling," in *Business Process Management: Models, Techniques and Empirical Studies*, W. van der Aalst, J. Desel, and A. Overweis, Eds. Berlin et al.: Springer, 2000, pp. 30–49.
- [22] W. Kamlah and P. Lorenzen, *Logische Propädeutik: Vorschule des vernünftigen Redens*, 3rd ed. Stuttgart: Metzler, 1996.
- [23] J. Mendling, J. Recker, and J. Wolf, "Collaboration Features in current BPM Tools," *EMISA Forum*, vol. 32, no. 1, pp. 48–65, 2012.
- [24] T. Gross, "Supporting Effortless Coordination: 25 Years of Awareness Research," *Comput. Support. Coop. Work*, vol. 22, no. 4–6, pp. 425–474, Jun. 2013.
- [25] C. Gutwin, R. Penner, and K. Schneider, "Group awareness in distributed software development," in Proceedings of the 2004 ACM conference on Computer supported cooperative work - CSCW '04, 2004, pp. 72–81.
- [26] H. Kagdi, M. L. Collard, and J. I. Maletic, "A survey and taxonomy of approaches for mining software repositories in the context of software evolution," pp. 77–131, 2007.
- [27] A. Chen, E. Chou, J. Wong, A. Y. Yao, and A. Michail, "CVSSearch: searching through source code using CVS comments," *Proc. IEEE Int. Conf. Softw. Maintenance. ICSM 2001*, pp. 364–373, 2001.
- [28] C. M. Neuwirth, D. S. Kaufer, R. Chandhok, and J. H. Morris, "Issues in the Design of Computer Support for and Commenting," in *CSCW '90*, 1990, no. October, pp. 183–195.
- [29] G. Decker and M. Weske, "Towards Collaborative Business Process Modeling," *Cut. IT J.*, 2009.
- [30] W. Tichy, "RCS—a system for version control," *Softw. Pract. Exp.*, vol. 7, no. July 1985, pp. 637–654, 1985.
- [31] Y. S. Maarek, D. M. Berry, and G. E. Kaiser, "An Information Retrieval Approach For Automatically Constructing Software Libraries," *IEEE Trans. Softw. Eng.*, vol. 17, no. 8, pp. 800–813, 1991.
- [32] N. Clever, J. Holler, J. Püster, and M. Shitkova, "Growing Trees – A Versioning Approach for Business Process Models based on Graph Theory," in Proceedings of the European Conference on Information Systems (ECIS) 2013, Paper 157, 2013.
- [33] E. J. Choi and Y. R. Kwon, "An efficient method for version control of a tree data structure," *Softw. Pract. Exp.*, vol. 27, no. 7, pp. 797–811, 1997.
- [34] R. Tarjan, "Depth-First Search and Linear Graph Algorithms," *SIAM J. Comput.*, vol. 1, no. 2, pp. 146–160, Jun. 1972.
- [35] C. E. Shannon, "A Mathematical Theory of Communication, Part I," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.
- [36] P. P.-S. Chen, "The Entity-Relationship Unified View of Data Model-Toward a," *ACM Trans. Database Syst.*, vol. 1, no. 1, pp. 9–36, 1976.

An Ontology for Formalizing and Automating the Strategic Planning Process

Juan Luis Dalmau-Espert, Faraón Llorens-Largo and Rafael Molina-Carmona
Group "Informática Industrial e Inteligencia Artificial"
University of Alicante
Alicante, Spain
e-mail: {jldalmau, faraon, rmolina}@dccia.ua.es

Abstract—From late 1980s and early twenty-first century, the environment where the organizations develop and act, has become increasingly uncertain and complex. Under these conditions, organizations have detected the need to move towards a more participatory model to address and reduce the complexity, based on information and knowledge as core assets to reduce environmental uncertainty and thereby ensure better decision-making. This new form of governance involves changes in the Strategic Planning process that are aligned with the characteristics of the new organizational model. Ontologies, as theories of content that allow the formalization of processes and knowledge, are a key element in this context. The aim of this paper is to formally define an ontology that could be defined in the future using the Web Ontology Language (OWL) that meets the standards approved by the World Wide Web Consortium (W3C) and that is used to formalize the process of SP, as well as the knowledge that is created and flows among the several participants in the process.

Keywords—ontologies; strategic planning; intelligent organizations; collective intelligence.

I. INTRODUCTION

From late 1980s and beginning of XXI century a series of events have occurred that, taken as a whole, paint a new landscape within the world of organizations in which the environment around them is increasingly uncertain and complex. Some of the most important facts behind this assertion are the internationalization of organizations, the economy globalization, the technological dynamism and, as a result, the growing need and importance of knowledge and learning within organizations [1].

Knowledge and ability to learn are today the main assets of the new model of intelligent organization referenced in [2]. The complexity and uncertainty of the environment can be reduced by providing and managing valuable information to help detect new needs in the environment and to guide the realization of new ideas to compete. To this end, the model of intelligent enterprise stands on three pillars: the collective intelligence, knowledge management and information and collaboration technologies.

The Collective Intelligence (CI) is the ability of an organization to engage its stakeholders in a task of intellectual cooperation to ask questions and find answers to questions concerning the organization [2].

The Knowledge Management was initially defined by Nonaka and Takeuchi [3] as the process of applying a systematic approach to the capture, structuring, management,

and dissemination of knowledge throughout an organization to work faster, reuse best practices, and reduce costly rework from project to project.

The information and collaboration technologies are the quantity and quality of software, hardware and networks facilitating relational and information flows.

From the new model of intelligent organization, it can be concluded that the traditional paradigm and process (centralized and reactive) for strategic planning that organizations utilize must be adapted to also be focused on that shared, participatory and collaborative learning-based vision, which supports the intelligent company [4]. The objective is to reduce complexity and increase the agility and flexibility of the process.

The process of Strategic Planning (SP) is defined as the process by which managers of the firm analyse the internal and external environments for the purpose of formulating strategies and allocating resources to develop a competitive advantage in an industry that allows for the successful achievement of organizational goals [5].

From the previous definition and the new needs of organizations, two key issues of formalization should be addressed and solved within the new Strategic Planning:

1) Formally define a conceptual framework that serves to represent the information/knowledge extracted from the internal and external environment of the organization and from each participant in the process, so that it constitutes a common vocabulary with which all managers can have a unified vision of the facts and that they can use to communicate and cooperate.

2) Formalize the Strategic Planning process itself to determine the steps that make up this process (listed in [6]), the information/knowledge type that is used and who is involved in each case.

The problem of formalization and conceptualization commonly appears in the literature linked to the concept of ontology. In 1993, Gruber [7] originally defined the notion of ontology as an "explicit specification of a conceptualization". In 1997, Borst [8] defined ontology as a "formal specification of a shared conceptualization". Finally, in 1998, Studer [9] merged these two definitions stating that: "An ontology is a formal, explicit specification of a shared conceptualization". These definitions identify two key aspects to keep in mind: on the one hand this formalization allows for a strict description of ambiguities-free knowledge that can be machine-readable and, on the other hand, it

reflects the idea of sharing knowledge among individuals in a group.

Guarino stated in [10] a classification of ontologies based on the level of generality proposing four types: high-level, domain, task and application ontologies. The latter describes concepts simultaneously belonging to a domain and a particular task.

In this paper, we present an application ontology whose domain covers all the terminology associated with the SP and its implied task/process.

According to Gruber [7], ontologies consist of concepts (formalized basic ideas obtained from the domain), relationships (they represent the interaction and the connection between the domain concepts), functions (they identify an element by calculating a function on several elements of the ontology), instances (they represent certain objects of a concept) and axioms (declarative theorems on relations that the elements of the ontology must fulfil).

There are several methodologies and languages to define ontologies. The methodologies proposed by Noy and McGuinness [11] and Uschold [12] and the languages proposed by Gruber [7] and Smith et al.[13] are considered as the most notable.

In Section I an overview about Strategic Planning inside organizations is introduced. The definition and steps of the strategic planning process are described in Section II. Ontologies are formally defined in Section III and finally, in Section IV, the ontology for strategic planning process is presented.

II. STRATEGIC PLANNING PROCESS

According to Hill and Jones [6], the Strategic Planning process is defined on the basis of a model that consists of five main steps:

- 1) Selection of the corporate mission and major corporate goals.
- 2) Analysis of the external competitive environment of the organization to identify opportunities and threats.
- 3) Analysis of the internal operating environment of the organization to identify strengths and weaknesses.
- 4) Selection of the strategies that are build on the strengths of the organization and correct their weaknesses to take advantage of external opportunities and oppose external threats.
- 5) Implementation of the strategy.

In practice, there are several strategic planning models which contain the steps of the previously presented model but that may have some special features. In Figure 1 a graphical representation of the model of SP stated by Llorens [14] is presented. It is possible to establish which element of the SP model is defined in each aforementioned step of the process [15].

Mission, vision, values and strategic axes are determined in step 1 and they define the reason for the organization, the state to be achieved in the future, the areas of action to

achieve the vision and the organizational culture, respectively.

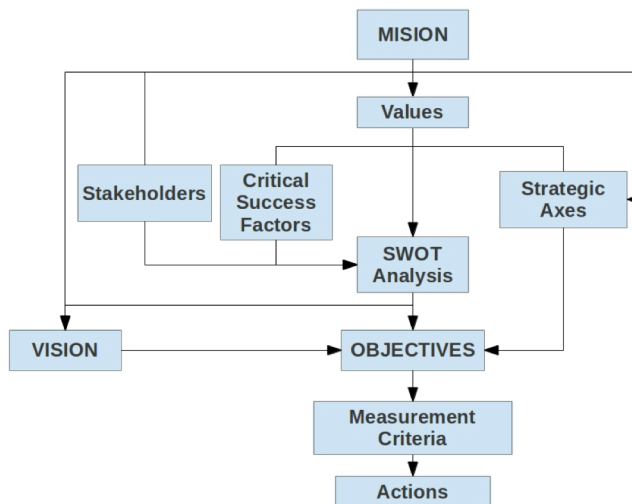


Figure 1. SP model, obtained from [14]

Critical Success Factors are determined in steps 2 and 3 from an analysis of the phenomena of the organization environment (internal and external) that can positively or negatively affect the fulfilment of the mission and a subsequent selection of the key factors.

The Strengths, Weaknesses, Opportunities and Threats Analysis (SWOT), the Objectives and the Measurement criteria are part of step 4. SWOT Analysis [16] is an analysis tool for decision-making. From the Critical Success Factors, the main strengths, weaknesses, threats and opportunities of the organization are identified. The following step is to determine which strengths and weaknesses are the most relevant to take advantage of the opportunities and to avoid the threats. The result is the Strategic Solution.

The Strategic Objectives are the major changes to perform so that the vision is achieved while fulfilling the mission. They should respond to the problem and to the identified strategic solution. At least one Strategic Objective must be defined for each Strategic Axe.

The Measurement Criteria are specific and usually quantitative targets for determining how far the organization is fulfilling its Strategic Objectives.

The Action Plans or Actions are a set of initiatives that are necessary to achieve the fulfilment of the Measurement Criteria and thus the Strategic Objectives.

The Stakeholders are those individuals, groups of individuals and institutions whose actions can positively or negatively influence the accomplishment of the Mission. They are important because, on one hand, they are involved in the phases of analysis and strategic selection and, on the other hand, the success of the strategies depends critically on the position and commitment of these individuals (internal and external) to bring them to fruition.

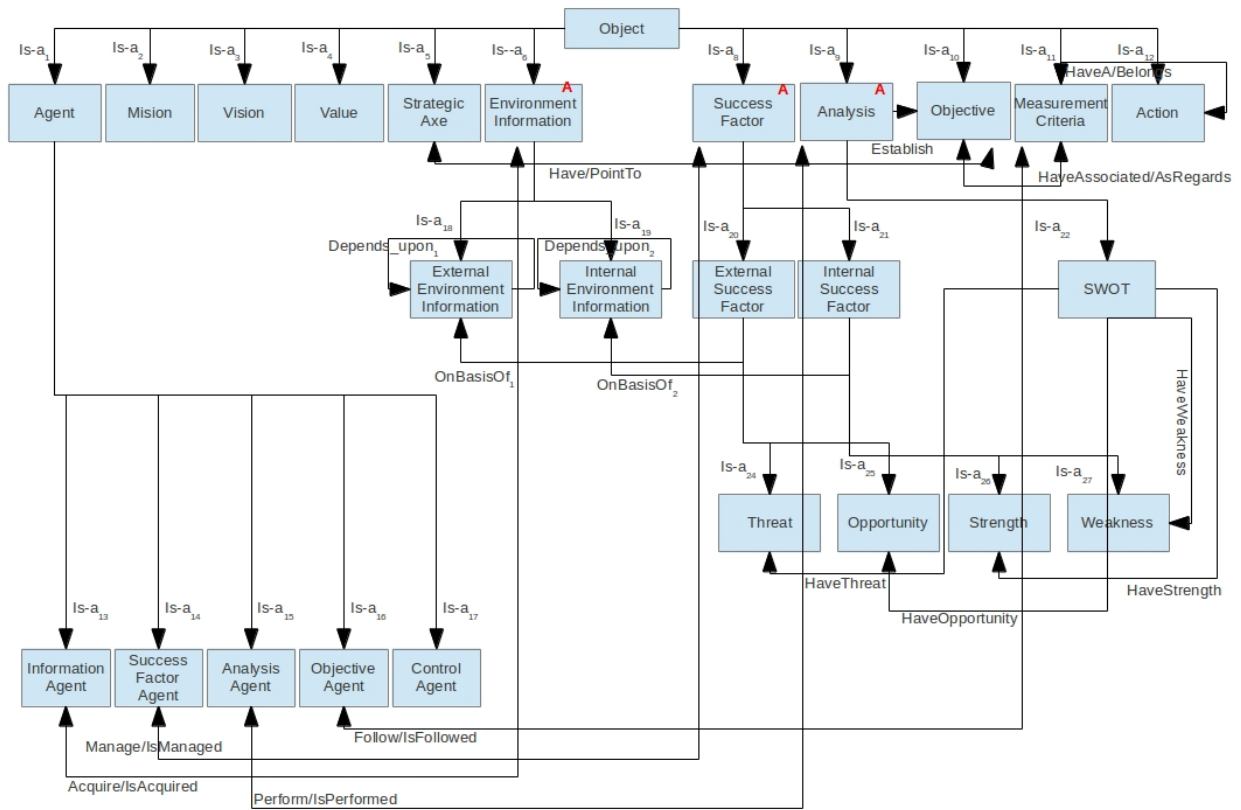


Figure 2. Ontology for SP process

Nowadays, the organizations approach the SP process using, in the best-case scenario, several tools and software technologies for some of the aforementioned steps, to achieve a certain degree of automation and formalization. This is the case of tools for Competitive Intelligence (CI) [17] or Business Intelligence (BI) [18] for the analysis of the environment. However, they are not integrated with the rest of the SP process and, often, they just provide reports and documents which substantially decrement the agility of the subsequent revisions of the Strategic Plan. There are approaches towards strategy content management [20] that consider the strategic planning process couldn't be automated. In this paper, an ontology is presented to achieve some degree of automation and parallelism in the strategic planning process.

III. ONTOLOGIES

According to Pretorius [19] the ontological structure (explicit specification of the conceptualization of the domain) is formally defined as expressed by (1).

$$O = \{C, R, A^o\} \quad (1)$$

where:

- C is a set of elements called concepts / classes, which have properties that describe their features and attributes.
- $R \subseteq C \times C$ is the set of relationships between the concepts / classes of C , which is defined so that it

contains the existing inherent hierarchical structure between concepts of C (hierarchical taxonomy).

- A^o is the set of axioms in O that impose restrictions on the concepts and their relationships.

The lexicon (language) L (common vocabulary regarding the conceptualization O) is defined as expressed by (2).

$$L = \{L^C, L^R, F, G\} \quad (2)$$

where:

- L^C is the set of elements called lexical entries of concepts.
- L^R is the set of elements called lexical entries of relationships.
- $F \subseteq L^C \times C$ is a reference to concepts that establishes the link between a concept and a lexical entry.
- $G \subseteq L^R \times R$ is a reference to relationships that establishes the link between a relationship and a lexical entry.

From the previous definitions an ontology O_m is formally defined as expressed by (3).

$$O_m = \langle O, L \rangle \quad (3)$$

Where O is the ontological structure and L is the corresponding associated lexicon (language).

IV. ONTOLOGY FOR STRATEGIC PLANNING PROCESS

An ontology O_m of type application is proposed. In Figure 2, the concepts and relationships contained by

structure O , as well as entries of lexicon L , which constitute the common vocabulary with which to refer to them, are shown. The Web Ontology Language (OWL) format of the ontology for SP process is not included for reasons of space.

The basis of the design of O_m is the Strategic Planning model of Figure 1. This design for ontology O_m includes:

- The formalization of all concepts associated with this model and involved in the process and their properties and relationships between them. An extract of dictionary of the associated concepts is included in Table I.
- The formalization of existing tasks/steps and the order to perform them in the SP process, obtained from the dependencies and relationships between the various concepts.
- The formalization of the Stakeholders as types of agents that are associated with a particular task or concept in the SP process.

The concepts are organized into levels based on a hierarchical structure (hierarchical taxonomy), which determines the inheritance of properties between a father concept (or class) and the child concept (or class). All classes directly or indirectly inherit Object properties.

For practical purposes, when performing the SP process to obtain a Strategic Plan, the definition of this ontology will allow instantiating the concepts that serve to define and formalize how the process is done. Since all concepts inherit from concept Object, any created instance has a property Identifier that uniquely identifies the instance, a property Name to refer to it and a property Level that determines the step in the SP process which this instance is associated to. For example, concepts Mission, Vision and Value belong to the first level, since they are needed to be instantiated in step 1 of the SP process. However, concepts SWOT and Analysis Agents belong to level 4, because they take part in step 4 of the process.

As a summary, considering the aforementioned and the steps that the SP process comprises and that have been listed above:

- The completion of the first step would involve creating an instance of the concept Mission, among others, in which the Description property would be informed so that the mission of the organization in the SP process is fixed.
- In steps 2 and 3, some instances should be created, to define what information is needed, where is it obtained and who is involved in it. In this case they are instances of the concepts External Environment Information, Internal Environment Information and Information Agent. In the instances of the first two concepts, it can be specified where the information is obtained by defining the property Source, the type of value to be obtained (property TypeValue), as well as who is responsible for obtaining the value of this information, provided by a reference to an instance of the concept Information Agent.

TABLE I. EXTRACT OF O_m CONCEPTS DICTIONARY

Super Concept	Concept	Properties	Type	Reference
	Object	Identifier	String	
		Name	String	
		Level	Integer	
Object	Mission	Description	String	
Object	Environ. Informat.	Type	String (Internal/ External)	
		Value	String/ Decimal	
		Source	String	
		TypeValue	String (Quantity/ Quality)	
		Calculation Method	String	
		IsAcquired	ReferenceTo	Information Agent
Environ. Informat.	External Environ. Informat.	OnBasisOf ₁	ReferenceTo	External Environ. Informat.
Environ. Informat.	Internal Environ. Informat.	OnBasisOf ₂	ReferenceTo	Internal Environ. Informat.
Object	Agent	Frequency	String	
		AgentType	String (Information /Objective,...)	
		Module	ReferenceTo	Program Code
Agent	Information Agent	Acquire	ReferenceTo	Environ. Informat.

From the last point, a first advantage can be identified: comparing to the way the task of analysing the environment using CI tools is traditionally performed, the use of the ontology directly integrates the diversity of information and knowledge within the SP process itself. Until now, it was necessary to use different CI tools depending on the type of information to be collected for later generating reports for the rest of the SP process.

In addition, once the way of formally define and specify (using an instance) the information or a variable of the environment is set up, it can be reused for subsequent SP processes or reviews.

Moreover, the ontology itself provides some level of automation in the collection and updating of information, thanks to the property Frequency of concept Agent. This property determines how often the Agent has to update the value of the information that is related. This is interesting because of the need of organizations to regularly make adjustments or revisions of the SP, which implies a partial review of the whole process.

Due to the fact that agents represent the Stakeholders involved in the SP process, they have a property Module that contains the heuristics that is applied to obtain the value of the information to which they relate. Thus, each agent is

responsible for providing to the process the information it knows.

The fact that instances of Environment Information contain a property Name, allows the creation of other instances of Environment Information that uses them and belongs to a higher level of abstraction (knowledge). For example, let Env_1 and Env_2 be two instances of the Environment Information concept, whose names are $Env_1.Name = Country_Competition_Index$ and $Env_2.Name = European_Competition_Index$ and they are computed by two instances of the Information Agent concept Ag_1 and Ag_2 . It is possible to create another instance $Env_3.Name = Ratio_Competition_Index$ of the Environment Information concept with a higher level of abstraction, which depends upon (relation depends_upon_x in O_m) the others and is computed by an instance of the Information Agent concept Ag_3 . This way, the ontology enables two important features: the possibility of interaction between agents using a common vocabulary and its use to cooperate.

The relationships between concepts formalize the order in which the tasks of the process are carried out, due to the fact that they establish the restrictions and dependencies between instances of two associated concepts at different levels of the process. For example, to create an instance of the SWOT concept of level 4 it is necessary to create in advance instances of concepts Threat, Opportunity, Strength and Weakness of which it depends according to the relationships HaveThreat, HaveOpportunity, etc., which are of level 2 and 3.

V. CONCLUSIONS AND FUTURE WORK

The complex and uncertain environment surrounding today organizations have imposed the need for a change in the organizational model. The intelligent organization model requires a new form of governance in which it is essential that the Strategic Planning process has a more participatory approach allowing an increased participation and interaction of the stakeholders in the procurement and management of information/knowledge, and based on it, in making decisions that serve to achieve the synchrony of the organization with this environment. This is the way to reduce the uncertainty and complexity of the environment due to the fact that, on one hand, a large amount of characterizing variables is obtained and, on the other hand, they are obtained in an easier and more agile way thanks to the participation of several specialists (Stakeholders).

The proposed ontology allows the formalization of the SP process by fixing its steps and its dependencies, the concepts of the domain that take part in the process, as well as their relationships and the Stakeholders involved in it. The formalization of the concepts provides a common vocabulary, which the Stakeholders can use to communicate and interact throughout the completion of the process. Finally, the fact that the development of the SP process is translated into a set of instances with a formal well-defined structure allows its reuse in other SP processes, as well as obtaining a complete documentation of the process in a more accessible format that is habitual nowadays. Moreover, it is even possible to automate certain parts of the SP process, so

that the instances generated by previous operations are machine-readable (property of the ontology).

The presented ontology for SP process is the first stage to propose a formal, automated and agile model for Strategic Planning. The very next step is defining the ontology using OWL to obtain the subsequent generation of the associated classes of a programming language. The aim of this hierarchy of classes is allowing the development of a graphical tool to design and graphically display the SP process. It could be possible by incorporating a panel of visible objects that represent instances of the classes and which could be accessed to specify their properties. This graphical tool will be the germ of an integral system for automated SP.

REFERENCES

- [1] J. Ventura López. *Strategic Analysis of the Company*, Ed. Paraninfo CENGAGE Learning, ISBN 978-84-9732-302-4, 2009.
- [2] O. Zara. *Managing Collective Intelligence: Toward a New Corporate Governance*, Ed. M2 Editions, ISBN: 978-291-62-6026-6, 2008.
- [3] I. Nonaka and H. Takeuchi. *The knowledge-creating Company: How Japanese Companies Create the Dynamics of Innovation*, Ed. Oxford University Press, ISBN 0-19-509269-4, 1995.
- [4] P. M. Senge. *The Fifth Discipline: The Art and Practice of the Learning Organization*, Doubleday/Currency, ISBN: 978-847-57-7351-3, 2006.
- [5] M. Z. Cox, J. Daspit, E. McLaughlin, and R. J. Jones. "Strategic Management: Is It an Academic Discipline?". III *Journal of Business Strategies*, Vol. 29 Issue 1, pp. 27-28, 2012.
- [6] C. W. L. Hill and G. R. Jones. *Strategic Management: An Integrated Approach*, 8th Edition, Ed. McGrawHill. ISBN: 978-970-10-7269-1, 2012.
- [7] T. R. Gruber. "A Translation Approach to Portable Ontologies". *Knowledge Acquisition*, 1993, pp. 199-220.
- [8] W. Borst. *Construction of Engineering Ontologies*. PhD thesis, Institute for Telematica and Information Technology, University of Twente, Enschede, The Netherlands, 1997.
- [9] R. Studer, V. Richard Benjamins, and F. Dieter. "Knowledge engineering: Principles and methods", 1998, pp. 161-198.
- [10] N. Guarino. "Formal Ontology in Information Systems". In *Proceedings of FOIS'98*, Trento, Italy, IOS Press, Amsterdam, 1998.
- [11] N. F. Noy and D. L. McGuinness. "Ontology Development 101: A Guide to Creating Your First Ontology", 2001.
- [12] M. Uschold. "Building Ontologies: towards a unified methodology". *Proceedings of Expert Systems '96*, the 16th Annual Conference of the British Computer Society Specialist Group on Expert Systems, Cambridge, 1996.
- [13] M. Smith, I. Horrocks, M. Hörtzsch, and B. Glimm. *OWL 2 Web Ontology Language: Conformance (Second Edition)*, 2012.
- [14] F. Llorens. *Strategic Plan of the University of Alicante (Horizonte 2012)* <http://web.ua.es/es/peua/horizonte-2012.html>, 2007. [Retrieved: December, 2014]
- [15] C. A. Olivera Rodríguez. "Strategic Exercise: Facilitator Guide". Matanzas, 2011.
- [16] N. Pahl and A. Richter. *Swot Analysis - Idea, Methodology and a Practical Approach*, Ed. Books on Demand, 2009.

- [17] J. F. Prescott and S. H. Miller. Proven Strategies in Competitive Intelligence: Lessons from the Trenches, Ed. Wiley, 2004.
- [18] C. Howson. Successful Business Intelligence: Unlock the Value of BI & Big Data, Ed. McGraw Hill Education, ISBN: 978-0071809184, 2013.
- [19] A. Johannes Pretorius. "Ontologies - Introduction and Overview". Semantic Technology and Applications Research Laboratory, Vrije Universiteit Brussel, Belgium, Adapted from: PRETORIUS, A.J., "Lexon Visualisation: Visualising Binary Fact Types in Ontology Bases", Chapter 2, Unpublished MSc Thesis, Brussels, Vrije Universiteit Brussel, 2004.
- [20] S. Paradies, S. Zillner and M. Skubacz. "Towards Collaborative Strategy Content Management using Ontologies". Workshop on Collaborative Construction, Management and Linking of Structured Knowledge. International Semantic Web Conference. ISSN 1613-0073. Washington D.C., 2009.

A Usability Evaluation Methodology of Digital Library

Luz A. Sánchez-Gálvez^{1,2}

¹Facultad de Ciencias de la Computación,
Benemérita Universidad Autónoma de Puebla,
72570, Puebla, México
sanchez.galvez@correo.buap.mx
luzsg@correo.ugr.es

Juan M. Fernández-Luna

²Departamento de Ciencias de la Computación e I.A., E.T.S.
de Ingeniería Informática, CITIC-UGR
Universidad de Granada,
18071 Granada, Spain
jmfluna@decsai.ugr.es

Abstract— Digital Libraries are information systems that allow managing and preserving digital resources, as well as providing access to them. The designing of these kinds of systems should always be completed having in mind the users. Accordingly, it is of outmost importance that the user interface presents a high level of usability. In this paper, a usability evaluation methodology of Digital Libraries is proposed; specifically for the Web site of Academic Digital Libraries. The methodology offers a evaluation instrument that collects the users' perceptions through four dimensions (effectiveness, efficiency, satisfaction and learnability). In addition, the gap theory of quality service is employed and a fuzzy linguistic approach using aggregation operators, which operate directly with words (linguistic information) is applied. Therefore, the methodology shows to be a significant, innovative contribution to the research area on usability evaluation of digital libraries. It can be useful for both an academic library Web site and an operational digital library. A case study is presented in order to demonstrate the usage of usability evaluation methodology of digital library.

Keywords-digital libraries; usability evaluation methodology; usability dimensions; evaluation instrument; aggregation operators.

I. INTRODUCTION

There are many definitions of usability in the literature, which provide different and complementary points of view, and show the evolution of the term itself along the evolution of knowledge. Usability is the broad discipline of applying sound scientific observation, measurement, and design principles to the creation and maintenance of Web sites in order to bring about the greatest ease of use [1]. According to Nielsen [2], usability focuses mainly on the use of a Web site or an interface and how people use it to carry out their tasks. If a Web site or interface is not able to satisfy the needs of their users, they will not be successful in the long term.

Usability has several international standard definitions. It is a quality attribute that assesses how simple it is to utilize user interfaces. The concept of usability, is also related to methods to improve the easiness of its use throughout the design process.

Digital Libraries (DLs) are information systems that allow to manage and preserve digital resources, and to provide access to them. The design of these systems must be always accomplished bearing in mind the user, so, it is

vitaly important that the user interface has a high degree of usability. In this paper, a usability evaluation methodology for the DLs —specifically for the academic DL Web site— such as a case study of the usability evaluation of the DL Web site in the University of Puebla is presented. The methodology proposes an evaluation instrument, that collects the users' perceptions by four dimensions (effectiveness, efficiency, satisfaction and learnability). In addition, the gap theory of service quality and a fuzzy linguistic model using aggregation operators of linguistic information —which operate directly with words— are used. Therefore, the methodology proves to be a significant and innovative contribution to the area of usability evaluation research of DLs. The same can be applied to a Web site of an academic DL as an operational DL.

The rest of this paper is organized as follows: Section II describes some related works in evaluating DLs. Section III explains the LibQUAL+ methodology. Section IV presents a brief introduction to the linguistic approach. Section V details a usability evaluation methodology. Section VI outlines the conclusions and future work.

II. RELATED WORK

According to the literature, most evaluations have focused on library service quality and its collections. Among the contributions of research on usability evaluation of Web sites, there are those of Hammill [3], who evaluated the usability of the Florida International University libraries' Web site by means of formal usability test and questionnaire, for determining whether its design and organization allow users to easily locate information based on the navigation, the clarity of vocabulary, and the visibility of the different sections. Oulanov and Pajarillo [4] described the results of a usability evaluation study of the Web-based graphical user interface version of the Web bibliographic database of the City University of New York, using questionnaire that determined the affect, efficiency, control, helpfulness, and adaptability as usability attributes. Lee [5] evaluated the usability of a Research Center library Web site in Korea by using a mixture of observation methods, and some formal usability tests; including heuristic evaluation, laboratory usability testing, and remote usability testing. Jeng's usability model [6], which is one of the most cited works on usability evaluation of DL Web sites, carried out usability tests based on tasks, comprising four usability dimensions; effectiveness, efficiency, satisfaction, and learnability, as

well as, some sub-attributes of usability. She also suggested some specific measures for each dimension, although it must be considered that those could vary according to the user's particular skills. Joo, Lin, and Lu [7] proposed a usability evaluation model by means a questionnaire, to survey the three usability criteria: effectiveness, efficiency, and learnability, in academic library Web sites. Alasem [8] used a questionnaire based usability test as a method for evaluating the usability of Saudi Digital Library's interface. In his usability evaluation, he suggested criteria such as efficiency, effectiveness, aesthetic appearance and learnability.

The usability study as in [9] is among the few studies comparing users' performance in multiple operational DLs: ACM, IEEE CS, NCSTRL, and NDLTD. This study used questionnaire and formal usability test, their objective was to identify specific characteristics that aid in the effectiveness (ease of use), likability, learnability and usefulness of DLs. The user's performance was measured by its ease of use, the amount of searching time, and the number of errors made. In [10] explored usability issues (perceived ease of search, satisfaction and perceived ease of browsing) on the design and the interaction in the browse and search function of ACM, IEEE CS, and IEEE Xplore.

LibQUAL+ methodology [11] was developed in the evaluating of the digital library services, provides a useful framework for usability evaluation of digital library. LibQUAL+ uses the gap theory of service quality, as well as other assessment frameworks based on SERVQUAL model [12]. Nowadays, The DigiQUAL project, being an extension of LibQUAL+, developed a service quality model reflecting digital environments [13].

Specialized assessment's works for academic libraries have been limited to assessing the quality service, but the evaluation of the Web site usability of academic libraries has received relatively little attention. Furthermore, these studies use neither the service quality gap nor the linguistic aggregation operators (LOWA and LWA). Regarding works that employ linguistic aggregation operators; these are strictly oriented to the quality of service [14] and they do not utilize the four usability attributes proposed hereby.

This paper provides a methodology that combines methods and principles —dimensions, questionnaire and measurement scale— of usability evaluation and fuzzy logic techniques, focusing on usability evaluation of digital libraries, which can either be academic or operational.

III. LIBQUAL+ METHODOLOGY

The Association of Research Libraries (ARL) in conjunction with the University of Texas A&M, started a project to obtain a standardized measure of the quality library service. The LibQUAL+ methodology is the result of such a project, which allows determining the quality of library services from the users' perception. LibQUAL+ is based on the theory of quality service —assessment applied in the environment of enterprises and organizations— particularly on the SERVQUAL evaluation methodology.

SERVQUAL is the most accepted and extended measurement of quality service. It is based on the principle

that "only customers judge quality; all other judgments are essentially irrelevant" [15]. Thus, customer satisfaction is the key element in SERVQUAL. Service quality is related to diminishing the distance between the customers' expectations and his final perception.

According to SERVQUAL, customers will evaluate, positively or negatively, the quality of a service where their prior perceptions were either higher or lower than expected. Hence, companies or organizations that provide services, where one of the objectives is being observed differently by means of a quality service, must show special interest on exceeding their customers' expectations.

LibQUAL+ looks forward to evaluating the quality of service of a library, considering three dimensions: the affective value provided by the staff, the value of the library space, and the value represented by the information control. Hence, service quality is evaluated through a survey of 22 questions; Minimum required level of service; Expected level of service; and the level perceived by the user. The Expected level and the Minimum required level establish the boundaries of a zone of tolerance within which the perceived scores should desirably float. Based on the users' feedback, it is possible to define two variables for detecting the strengths and weaknesses of a library i.e. Adequacy of Service (the difference between the perceived value and the minimum value) indicating the areas where the library service is below the level expected by the user, and Service Excellence (the difference between the perceived value and the expected value) that identifies areas where the library provides a better service than that expected by the user.

IV. LINGUISTIC APPROACH

The information cannot always be evaluated in a quantitative manner, sometimes it is necessary to do it qualitatively. The existence of qualitative variables inherent to human behavior, or external environment elements, which are difficult to quantify objectively lead individuals to express their opinions better, by using linguistic terms instead of precise numerical values. A linguistic variable differs from a numerical one in that its values are not numbers, but words or sentences in a natural or artificial language [16].

When a linguistic model is used, the existence of a suitable set of terms or labels according to the problem domain is assumed, then, the individuals can express their perceptions. The ordinal fuzzy linguistic model [17] is very useful as it simplifies the computing by eliminating the complexity of having to define a grammar.

An ordinal fuzzy linguistic modeling [16] is used in this paper to represent the users' perceptions with words, based on the linguistic aggregation operators LOWA and LWA [14], in order to evaluate the academic digital libraries Web sites usability.

The Linguistic Ordered Weighted Averaging (LOWA) is an operator used to aggregate non-weighted ordinal linguistic information, i.e., linguistic information values with equal importance. The Linguistic Weighted Averaging (LWA) is an operator used to aggregate weighted linguistic information, i.e., linguistic information values has different

importance. In order to calculate both operators, this paper follow the definitions established on [18].

V. A USABILITY EVALUATION METHODOLOGY

In this work, a usability evaluation methodology has been developed for the academic DL Web sites based on literature review. Therefore, this methodology takes an approach that requires the establishment of dimensions, as a way to measure usability, based on standards such as ISO 9241-11 [18] and Nielsen's definition [2]. Consequently, it considers four dimensions: effectiveness, efficiency, satisfaction, and learnability represented in twenty items to capture the users' perceptions to assess the usability degree of academic DL Web sites. Furthermore, the methodology uses a fuzzy linguistic model by means of aggregation operators with linguistic information, which handle words directly. They are important to allow sorting and classifying all data from an aggregation process, without any loss of linguistic information. In addition, the model LibQUAL+, which attempts to measure the overall service quality in academic DL is utilized. LibQUAL+ emerged from the SERVQUAL methodology; an instrument based on the gap theory of service quality, which was used to assess private sector institutions.

Deficiencies on usability of DL Web sites could be identified through LibQUAL+. Therefore, the proposed methodology of usability evaluation in this paper, could offer commendations for prioritizing improvements and guaranteeing a proper interface design of DL, based on users' preferences within an institution.

The usability evaluation methodology, consists of different steps; the development, production, implementation, evaluation and reliability of the questionnaire, as well as the results analysis, from where a series of recommendations to improve the usability of DLs Web sites can be provided. The steps for implementing this methodology are:

- A. The identification of dimensions.
- B. The preparation of the questionnaire.
- C. The usability evaluation by aggregating operators.
- D. The presentation of evaluation results.
- E. The commendation of the DL Web site.

A. Dimensions of Usability

The usability is a multidimensional concept. In this paper four dimensions of usability —effectiveness, efficiency, satisfaction and learnability— are proposed to assess the usability of academic digital libraries' Web sites (see Table 1).

In order to define the usability dimensions, the models of Nielsen [19], ISO 9241-11 [18], Shackel [20] Tsakonas [21], Jeng [6], and Xie [22] were revised. Finally, the chosen dimensions were based on the standard definition of ISO 9241-11 [18] and the Nielsen model [19]. Nielsen's model, which is one of the most cited in the area of usability engineering, postulates five attributes: learnability; efficiency; memorization; low error rate (easy error recovery); and subjective satisfaction.

TABLE I. DIMENSIONS OF METHODOLOGY

Dimension	Sub criteria	Definition
Effectiveness		It refers to completion of tasks where users achieve specific goals.
Efficiency		It refers to the resources used for performing a task.
Learnability		The system should be easy to learn and to understand; it should be easy for the user to achieve a task by using the system.
Satisfaction	Ease of use	It refers to the user's perception about the use of the system.
	Information Organization	It is to assess whether the structure, design, and organization of the system reach the users' goals.
	Clear Labeling	It refers to the clear labeling of the DL Web site from the users' point of view, and whether the terminology used is easy to understand.
	Visual Aspect	It evaluates the site design concerning its visual appealing.
	Error Recovery	It has to do with the easy to recover from errors made by the users.
	Navigability	It refers to the easiness that users may have to go from one site to another .

B. Questionnaire

An important part of the usability research has been the designing of a questionnaire. The items for measuring the usability of a DL Web site, were based on the literature on usability evaluation studies. First, it was necessary to establish the dimensions. To be able to generate measurement items; usability frameworks, usability guidelines, and usability testing were reviewed [6][11][19]. All measurement items chosen, were modified to reflect the unique features of academic libraries Web sites. Thus, twenty items establish the questionnaire (see Table 2) to capture the users' perceptions on the rate of usability of academic DL Web sites, based on the proposed usability dimensions. Users must answer the questions about their personal experience when interacting with the DL Web site.

C. Evaluation

Before evaluating the usability of the DL Web site of the University of Puebla —which is the case study—, the participants were asked to fill out a pre-questionnaire concerning demographic data —level of education, age, and gender, use frequency of the DL Web site and level of computer skills. The filling of the pre-questionnaire and usability questionnaire, were accomplished by accessing <http://encuiti.netai.net/>. In this case, it is a tool designed and implemented especially for this evaluation study; that allows the capture and analysis of data, by using the aggregation operators for data processing through LOWA and LWA.

A total of 54 users participated in the usability evaluation of the DL Web site of the University of Puebla —including students and teachers, both undergraduate and masters.

The pre-questionnaire was analyzed using frequency charts to determine the type of user, who responded to the survey based on their age, sex, level of education, computer

skills, and Web site usage. The usability evaluation questionnaire was examined bearing in mind the following steps:

TABLE II. QUESTIONNAIRE

Dimension	Item
Effectiveness	Can you usually complete a search task by using the DL Web site?
	Can you successfully find digital resources through the DL Web site?
	Do the digital resources of the DL Web site satisfy your needs of information?
	In general, is the DL Web site useful to help you find the information you are looking for?
Efficiency	Can you complete a task quickly by using the resources of the DL Web site?
	Do you obtain results quickly by using the DL Web site?
	Does the user interface gives you expeditious access to DL resources?
	Is the access to information services (databases, catalogs, etc.) quick and easy to use?
Learnability	Was learning to use the DL Web site easy?
	Are the terms used on the DL Web site easily understandable?
	Is the DL Web site help well organized?
	Are new users able to utilize the services without considerable effort?
Satisfaction	What is your main fulfillment when using the DL Web site?
	What is the rate of navigability of the DL Web site?
	Is the organization and distribution of information on the DL Web site clear?
	Is the language used on the Web site, appropriate and classified by tags clearly enough?
	Is the Web site visually appealing?
	Does the Web site allow an easy error recovery?
	Are he services offered by the DL Web site satisfactory?
Total Usability	What is the general usability of the the DL Web site?

1. The users expressed their judgment by completing the questionnaires (see Table 1).

The scale for this model is $S=\{VL=Very\ Low, L=Low, M=Medium, H=High, VH=Very\ High\}$.

As a result for each one of the users $u_j \in (u_1, u_2, \dots, u_n)$ and for each questionnaire item $i_k \in (i_1, i_2, \dots, i_m)$, m is the total number of questions; there is a tuple $(mv_{jk}, pv_{jk}, ev_{jk})$ of the minimum value — mv —, perceived value — pv — and the expected value — ev —, for each user u_j and for each question i_k .

2. To compute the global users' opinion concerning each item i_k of the tuple $(mv_{jk}, pv_{jk}, ev_{jk})$, the following aggregation operators are used:

2.1 LOWA [14] is used if all users are considered to bear the same importance.

$$\begin{aligned} mv_k &= \Phi_Q(mv_{1k}, \dots, mv_{nk}) \\ pv_k &= \Phi_Q(pv_{1k}, \dots, pv_{nk}) \\ ev_k &= \Phi_Q(ev_{1k}, \dots, ev_{nk}) \end{aligned} \quad (1)$$

2.2 LWA [14] is used when each user is considered to bear a different level of importance.

$$mv_k = \Phi_Q((UI(u_1, i_k), mv_{1k}), \dots, (UI(u_n, i_k), mv_{nk})) \quad (2)$$

$$pv_k = \Phi_Q((UI(u_1, i_k), pv_{1k}), \dots, (UI(u_n, i_k), pv_{nk}))$$

$$ev_k = \Phi_Q((UI(u_1, i_k), ev_{1k}), \dots, (UI(u_n, i_k), ev_{nk}))$$

Where $UI(u_j, i_k) \in S$ is the level of relative linguistic importance assigned to a user u_j for the item i_k .

3. The overall review of all questions of the tuple (mv, pv, ev) is calculated similarly to the previous step, by using aggregation operators:

LOWA [14] is used when all the items are considered to bear the same importance.

$$\begin{aligned} mv_k &= \Phi_Q(mv_1, \dots, mv_m) \\ pv_k &= \Phi_Q(pv_1, \dots, pv_m) \\ ev_k &= \Phi_Q(ev_1, \dots, ev_m) \end{aligned} \quad (3)$$

LWA [14] is used when each item is considered to carry a different level of importance.

$$\begin{aligned} mv &= \Phi_Q((II(i_1), mv_1), \dots, (II(i_m), mv_m)) \\ pv &= \Phi_Q((II(i_1), pv_1), \dots, (II(i_m), pv_m)) \\ ev &= \Phi_Q((II(i_1), ev_1), \dots, (II(i_m), ev_m)) \end{aligned} \quad (4)$$

Where $II(i_k) \in S$ is the level of relative linguistic importance assigned to item i_k .

4. The gap theory of service quality is applied to each item. The tolerance zone is located between the minimum and the expected values. The difference between the perceived and the minimum values, is called Service Adequacy —SA— and the Service Superiority —SS— is the difference between the expected values and the perceived ones. Therefore, for each item i_k , SA_k and SS_k are computed as follows [9]:

$$\begin{aligned} SA_k &= D(pv_k, mv_k) \\ SS_k &= D(ev_k, ev_k) \end{aligned} \quad (5)$$

On the other hand, when utilizing questionnaires for evaluating the usability of a Web site, it is important to verify the reliability of the evaluation instrument it is advisable to use the Cronbach's alpha.

Cronbach's alpha allows to quantify the level of reliability of a evaluation scale, built from k variables observed. Assuming that the variables are related to the qualitative interest data; the k variables should achieve stable, consistent measurements with a high level of correlation among themselves. A questionnaire is considered reliable when Cronbach's alpha is greater than 0.80. The formula for Cronbach's alpha is:

$$\alpha = \left[\frac{k}{k-1} \right] \left[1 - \frac{\sum_{i=1}^k S_i^2}{S_t^2} \right] \quad (6)$$

Where S_i^2 is the item variance i ;

S_t^2 is the item variance of all observed values;
 K in the item number of the questionnaire;

D. Results of the Questionaries

Both a quantitative and a qualitative analysis are accomplished in the usability evaluation of the DL Web site. The qualitative analysis focuses on calculating the aggregation operators LOWA and LWA; It is based on proposed linguistic labels on the scale.

The LOWA operator requires to obtain the combination of the users' perception for each item. Thus, Table 3 summarizes the result of the combined aggregation of the users' perception for the three assessed values: minimum,

perceived, and expected values, regarding the DL Web site as well as their corresponding gap. On the other hand, Figure 1 shows a radar chart that summarizes the user responses to the questionnaire items on the minimum, perceived and expected levels. This type of chart was used to display the results obtained with the LOWA operator. As shown therein, the usability of the minimum value is reflected on the chart with a medium value (orange color), by most users the perceived one (green color) and the expected one (blue color) show a tendency towards a higher level of usability in the DL Web site.

The LWA operator allows perceiving the opinion of all users on items with a different level of importance; which is suitable to evaluate the usability on this paper, because it contemplates four dimensions: efficiency, effectiveness, learnability, and satisfaction. So, the level of importance will vary according to the dimension being assessed. In case of measuring effectiveness, items 1, 2, 3, and 4 would have a Very High (VH) level of importance, while the remaining items present a Very Low (VL) level of importance as shown in Table 4.

TABLE III. RESULTS OF THE LOWA OPERATOR

Item	Minimum	Perceived	Expected	Usability Adequacy	Usability Excellence
1	M	H	H	L	VL
2	M	H	H	L	VL
3	M	H	H	L	VL
4	M	H	H	L	VL
5	M	H	H	L	VL
6	M	H	H	L	VL
7	M	H	H	L	VL
8	M	H	H	L	VL
9	M	H	H	L	VL
10	L	H	H	M	VL
11	L	H	H	M	VL
12	L	H	M	M	L+
13	L	H	M	M	L+
14	L	H	M	M	L+
15	M	H	H	B	VL
16	M	H	H	B	VL
17	L	H	H	M	VL
18	L	H	H	M	VL
19	M	H	H	L	VL
20	M	H	H	L	VL



Figure 1. Radial chart of the LOWA operator.

TABLE IV. LWA OPERATOR FOR THE EFFECTIVENESS DIMENSION

LWA: Effectiveness		
Minimum	Perceived	Expected
M	H	H

On the other hand, concerning the quantitative analysis, item 20 has been planted to measure the satisfaction of the user’s overall usability of the DL Web site. Figure 2 displays that 21 out of the 54 respondents have a High (H) overall satisfaction when evaluating the DL Web site usability; while 12 of them show a VH level; 16 show a Medium (M) value, and the other 5 present a Low (L) value.

As mentioned above, the evaluation questionnaire reliability, was calculated using Cronbach's alpha, obtaining a value equal to 0.91 for the minimum value; 0.91 for perceived value; and 0.92 for the expected value, which means such a reliability is fairly acceptable.

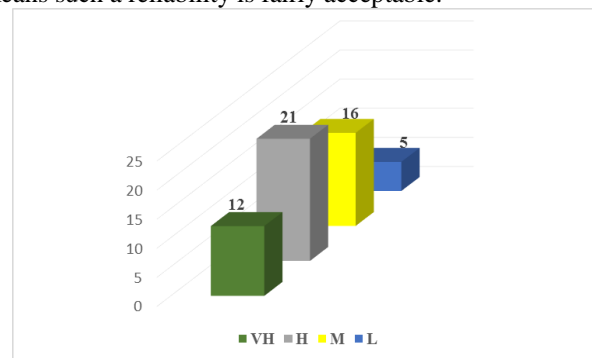


Figure 2. Chart of the Total Usability (20 items).

E. Commendations

On the whole, the usability evaluation results of the DL Web site in the University of Puebla, have been satisfactory. However, the adequacy gap indicates that the improvement should focus primarily on two-dimensions; learnability and satisfaction.

In the former, the adjustments should be directed to improve the functions of help. So, it is advisable to follow the guidelines provided by ISO, IBM and Microsoft, in order to meet the standards of the interface design. Thus, improving, incorporating, and maintaining the support functions visible (in addition to video tutorials) for guiding users —especially the novices—.

While in the latter, the changes should be oriented to improve the navigability, the interface and the error recovery. Consequently, including icons or links that allow return to a previous state of the system or even the main menu of the library is recommended as well as messages, which clearly inform and orient about the task of search being performed. As for the interface, three commendations are made: To modify its organization to simplify and improve its navigability; to focus on the services offered by the library, eliminating or reducing those that are strange to it, and using a clear terminology that would improve the appearance of the Web site. These interface adjustments help to simplify the management system errors. All the

aforementioned would facilitate the usability of the DL Web site in the University of Puebla, as a result, an excellent service can be provided.

VI. CONCLUSIONS AND FUTURE WORK

In this work, an innovative methodology for evaluating the usability of digital libraries has been developed. In developing the methodology, the basic principles of usability (establishing four dimensions, the questionnaire and the measurement scale) have been combined with models of service quality (the gap theory of service quality) and fuzzy logic models (LOWA and LWA linguistic operators) specifically applied in evaluating the usability of DLs.

This methodology could be used to evaluate any Web sites. However, some particular questions (such as responsive design and real time responses) should be analyzed, which are out of reach on this analysis.

In this research a measuring instrument (questionnaire) with 20 questions that collects user perceptions based on the four dimensions proposed to evaluate the usability of DL is intended. Moreover, the use of aggregation operators of linguistic information with a measurement scale was raised five linguistic labels. The gap in quality service has set the pace to suggest a number of commendations for improving the BUAP DL Web site.

Cronbach's alpha was used for verifying the reliability of the measurement instrument, resulting in a value close to 1, indicating a rather acceptable measurement instrument. A qualitative study was also carried out using descriptive statistics to compare the results with those obtained with the aggregation operators, which showed that the use of these operators is appropriate for the methodology.

Future work will focus on applying the survey again after a trial using the DL Web site in order to accomplish a comparative analysis, improving the survey tool used in such a manner that allows diverse linguistic quantifiers be used to calculate the weights of aggregation operators.

ACKNOWLEDGMENT

This paper has been supported by the Spanish "Ministerio de Ciencia e Innovación" under projects TIN2011-28538-C02-02 and TIN2013-42741-P.

REFERENCES

- [1] M. Pearrow, *Web site usability handbook*. Rockland, MA: Charles River Media. 2000.
- [2] J. Nielsen, *Usability engineering*. Cambridge, Mass: Academic Pr. 1993.
- [3] S. Hammill, "Usability testing at Florida International University Libraries: what we learned," *Electronic Journal of Academic and Special Librarianship*, vol. 4(1), 2003.
- [4] A. Oulanov and E.F.Y. Pajarillo, "CUNY+ Web: usability study of the Web-based GUI version of the bibliographic database of the City University of New York (CUNY)," *The Electronic Library*, vol. 20 (6), pp.481-487, 2002.
- [5] K. P. Lee, *A study on the improvement plan by analyzing user interaction pattern with the RISS*. Technical Report KR2004-17, KERIS, Seoul 2004.
- [6] J. Jeng, "Usability assessment of academic digital libraries: Effectiveness, efficiency, satisfaction, and learnability," *International Journal of Libraries and Information Services*, vol. 55, pp. 96-121. 2005.
- [7] S. Joo, S. Lin, and K. Lu, "A Usability Evaluation Model for Academic Library Websites: Efficiency, Effectiveness and Learnability," *Journal of Library and Information Studies* vol. 9 (2), pp.11-26, 2011.
- [8] A. Alasem, "Evaluating the Usability of Saudi Digital Library's Interface (SDL)," In *Proceedings of the World Congress on Engineering and Computer Science 2013 Vol I WCECS 2013*, San Francisco, USA, pp. 23-25, October 2013,
- [9] R. Kengeri, D. Seals, H. Reddy, H. Harley, and E. Fox, "Usability study of digital libraries: ACM, IEEE-CS, NCSTRL, NDLTD," *International Journal on Digital Libraries* Vol. 2, pp. 157-69, 1999.
- [10] X. Zhang, J. Liu, Y. Li, and Y. Zhang "How usable are operational digital libraries – A usability evaluation of system interactions. *Proceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, Pittsburgh, pp. 177-186, July 2009.
- [11] C. Cook, F. Heath, and B. Thompson, *Score norms for improving library service quality: A LibQUAL+ study*. *Portal: Libraries and the Academy*, vol. 2 (1), pp. 13-26, 2002.
- [12] A. Parasuraman, V.A. Zeithaml, and L.L. Berry, "SERVQUAL: A Multiple-Item Scale for Measuring Customer Perceptions of Service Quality," *Journal of Retailing*, vol. 64, pp.12-40, 1988.
- [13] M. Kyriallidou and S. Giersch, "Developing the DigiQUAL protocol for digital library evaluation," In M. Marlino, T. Summer, & F. Shipman (Eds.), *Proceedings of the Fifth ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 172-173. New York: ACM Press. 2005.
- [14] R. Heradio, F.J. Cabrerizo, D. Fernández-Amorós, M. Herrera, and E. Herrera-Viedma, "A fuzzy linguistic model to evaluate the quality of Library 2.0 functionalities," *International Journal of Information Management*, vol. 33, pp. 642- 654, 2013.
- [15] V.A. Zeithaml, L.L. Berry, and A. Parasuraman, *Delivering Quality Services - Balancing Customer Perceptions and Expectations*. The Free Press, New York, 1990.
- [16] L.A. Zadeh, "The concept of a linguistic variable and its applications to approximate reasoning. Part III," *Information Sciences*, vol. 9 (1), pp. 43-80, 1975.
- [17] F. Herrera, E. Herrera-Viedma, and J.L. Verdegay, "Direct approach processes in group decision making using linguistic OWA operators," *Fuzzy Sets and Systems*, vol. 79, pp. 175-190, 1996.
- [18] International Organization for Standardization, Technical Committee of Ergonomics, 1998. *Ergonomic requirements for office work with visual display terminals (VDTs): Part 11: Guidance on usability (ISO No. 9241-11)*.
- [19] J. Nielsen and R.L. Mack, *Usability inspection methods*, eds. 1994. New York: Wiley.
- [20] B. Shackel, *Usability - context, framework, definition, design and evaluation*. In B. Shackel & S. Richardson (Eds.), *Human factors for informatics usability*, pp.21-37. Cambridge: Cambridge University Press. 1991.
- [21] G. Tsakonas, S. Kapidakis, and C. Papatheodorou, "Evaluation of user interaction in digital libraries," *Notes of the DELOS WP7 Workshop on the Evaluation of Digital Libraries*. pp. 45-60, 2004.
- [22] H. Xie, "Users' evaluation of digital libraries (DLs): their uses, their criteria, and their assessment," *Information Processing and Management*. Vol. 44(3), pp. 1346-1373. 2008.

LOM, a Locally Oriented Metric which Improves Accuracy in Classification Problems

Julio Revilla Ocejó, Evaristo Kahoraho Bukubiye
 Dpt. of Industrial Technologies
 University of Deusto
 Bilbao, Spain
 jrevilla@deusto.es, kahoraho@deusto.es

Abstract— New tools for computer automatic reasoning, case-based reasoning or data mining, require powerful artificial intelligence techniques to provide both, a high rate of precision in their predictions, and a collection of similar past experiences that could be applied in the actual scenario. Algorithms based on Nearest Neighbors, Support Vector Machines, etc. often provide an accurate solution for classification problems, but they depend on how the similarity is measured. For this task, most of the experts employ a Euclidean metric that equally weights all attributes of the case (which is unlikely in the real world). In addition, it is well known that a correct metric choice should improve their prediction abilities and avoid the curse of dimensionality. In this paper, we present a new metric for those algorithms. It replaces the traditional Euclidean approach with a new riemannian metric that “enlarges” the space parallel to the frontier of separation between classes, thus improving classification accuracy.

Keywords- metrics; *k*-NN; SVM; Riemannian; Dijkstra.

I. INTRODUCTION

In a classification problem [1], a “case” is characterized by a set of numerical, ordinal and/or nominal values formed by its P attributes. It belongs to one of the J possible classes $\{y_j\}$. Classification algorithms based on empirical data have at their disposal N “training cases”, which consist of values of the attributes and their classes $\{x_n, y_n\}$, $n = 1 \dots N$. The target of these algorithms is assigning a new case to the correct class.

In current case-based reasoning (CBR), data mining tools (DM), etc. nearest neighbor classification algorithms (*k*-NN) are frequently used to implement the similar cases search phase. This kind of algorithms seeks for the cases closest to the new one to determine its class by majority voting. Internally, they usually employ a Euclidean metric to measure the “distance” (dissimilarity) among cases, but this metric does not behave particularly well in the border of separation between two classes [2][3][4].

We introduce, in this paper, a better metric for these algorithms, the LOM metric. It attempts to adjust locally the perception of which points are close to a given one. It adapts the measurement of distances in such a way that they are shorter in the direction parallel to the tangent hyperplane to the border of separation between classes, and gets larger in the perpendicular direction. The decision function, which provides the border of separation between classes, is estimated in a pilot trial and obtained by a traditional classification algorithm. LOM has only two free parameters, which can be optimally tuned through a cross-validation (CV) procedure.

This paper is organized as follows: Section 2 relates actual and past research in this area. Section 3 describes the LOM metric and its properties. In Sections 4 and 5, the specifics of the LOM metric, working in conjunction with a support vector machine (SVM) decision function, are explained. Section 6 contains the first results, which prove the merit of the algorithm. In the end, a section of conclusions and the work to be addressed in a near future is included.

II. METRICS USED IN SIMILARITY MEASUREMENTS

The vast majority of the algorithms used in automated reasoning based on previous cases, sensor fusion, DM and other typical tasks of artificial intelligence (AI), in one way or another, base its calculations on some kind of measure of similarity/dissimilarity between objects. In most scenarios, this fact usually remains unnoticed (since is regularly used the Euclidean metric as a default). Euclidean metric considers of equal relevance the values of the different attributes. It ignores in what area of the attributes space the case is located and whether the cases close to it belong or not to the same class.

To distinguish between similar objects, humans weight some attributes more than others; the features chosen to classify an object (and their relevance) depend on what they see at first glance. Definitively, humans do not employ a Euclidean metric.

A. Relevant previous studies on similarity metrics

Numerous studies have suggested that in the vicinity of the separation surface of two classes, equidistant distance curves (isolines) “should be enlarged” in the direction tangent to the surface and “shortened” in the perpendicular to it. Far away from this border, the Euclidean metric can be considered a sufficiently good choice.

The algorithm LAMANNA of Carlotta Domeniconi and her team [2][3], proposes to employ the boundary of separation between classes, provided by a SVM, to determine the most relevant local directions in the vicinity of a point. This paper introduces very interesting aspects, as using the gradient of the decision function as an indicator of the direction of greater relevance in the classification. However, their algorithm does not lead to a metric, since the distance does not meet properties, such as the triangular inequality or the symmetry ones. Neither is it clear that their distance will lead to positive definite (PD) kernels and, therefore, it is not guaranteed that it could be used by the common optimization algorithms employed in the SVMs.

Weinberger et al. [5] optimize a Mahalanobis metric using semidefinite programming (SDP). Their algorithm provides

acceptable results and leads to a global metric for the entire attributes space. It does also optimize a lot of parameters (P^2) so the final metric does not have easily interpretable properties. In similar approaches, other researchers [6][7][8] have attempted to optimize the Mahalanobis matrix using different techniques (SDP to optimize the separation between pairs of cases, average of weighted covariance matrices, etc.).

Goldberg et al. [9] proposed the NCA algorithm that adjusts the elements of a linear transformation matrix, minimizing the probability of classification error in a random selection of cases close to the current one. It results in a global metric, but the minimization method is subject to be trapped in local minima; and can incur in overfitting. In general, the use of metrics which involve matrices with real elements does not allow their main directions to change in different regions of the space of attributes, just as we propose in this paper.

Following a completely different approach, several authors have tried to improve the metric used to calculate the distances in SVMs based on RBF kernels. Chan et al. [10] proposed to modify the radius of the RBF depending on the density of cases in that region. It is a very simple approach and does not orient the metric in any special direction.

Amari and Wu [11][12] were pioneers in pointing out the relevance of the kernel choice in SVM performance. They proposed to use conformal transformations for RBF kernels, which widen the spatial resolution in the vicinity of the surface of separation between classes. Wu et al. [13] developed a conformal transformation in the space of the features, based on the support vectors (SV) obtained in a previous iteration.

Williams et al. [14] chose a different function to implement the conformal transformation depending on the value of the separation function provided by a SVM (whose value for different points in the attributes space provides the criteria for class separation).

All conformal transformation previously related, use a prior calculation of the border of separation between classes and then "widen" the metric in the associated Riemann space. None of them justify in detail why this would improve the separability of the points in those areas. They neither orient the metric in the direction of class separation.

III. THE PROPOSED METRIC

What does that a point is "close" to another mean? This is a fundamental question to define a metric. In a two-class problem, in the vicinity of a point in the attributes space, some directions point to areas where there are many elements of a certain class, while other directions point to areas with elements of mixed classes. To determine these directions is of fundamental relevance when delineating a flawless metric.

The aim of our actual research is to define a metric that possess two groups of directions in the P -dimensional attributes space:

- One direction is perpendicular to the hypersurface separating both classes. The gradient of the separation function could be a valid method to determine it.
- The set of all directions perpendicular to the former, i.e., all those contained in the hyperplane perpendicular to the gradient.

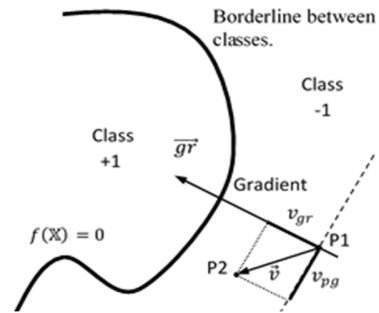


Figure 1. Main directions for the metric.

Different weights can be assigned to each of these two groups of directions. We propose to calculate the "distance" between two very close points, based on a weighted sum of the squared norms of the projections of the segment joining the two points on each of these two sets of directions. Hence, the formula proposed to calculate the distance is:

$$d^2 = \frac{v_{gr}^2}{r_m^2} + \frac{v_{pg}^2}{r_M^2} \tag{1}$$

where:

- \vec{v} is the vector joining the two points (P1-P2 in Fig.1) whose distance is to be evaluated.
- \vec{gr} is the normalized gradient vector at point P1.
- v_{gr} is the norm of the projection of \vec{v} on the direction of the gradient of the hypersurface of class separation. It can be easily computed by $v_{gr} = \vec{v} \cdot \vec{gr}$.
- v_{pg} is the norm of the projection of \vec{v} on the hyperplane perpendicular to the gradient. $v_{pg}^2 = \|\vec{v}\|^2 - v_{gr}^2$.
- r_m is the "minor radius". It reflects that in the direction of the gradient, the length of the projection will contribute to a larger distance between points.
- r_M is the "major radius". It weights the projection in the plane perpendicular to the gradient. Any separation in this direction will contribute to a lesser extent to the distance between points.

In order to relate r_m and r_M to the separation function, they will be defined by the following formulas:

$$\begin{aligned} r_m &= r / (1 + \tau e^{-f(\mathbf{X})^2}) \\ r_M &= r \cdot (1 + \tau e^{-f(\mathbf{X})^2}) \end{aligned} \tag{2}$$

where:

- r defines the general scale for the metric.
- τ is a parameter which shrinks/amplifies the minor/major radius as the base point approaches to the border of separation between classes.
- $f(\mathbf{X})$ is the function that evaluates the membership of a point \mathbf{X} in the attribute's space to one or another class. On the boundary between classes, its value is zero and as the point moves away, takes positive or negative values.

When the base point is close to the border of separation between classes, $f(\mathbf{X})$ is zero, and therefore, the major radius will be $(1 + \tau)^2$ times the minor radius. In those areas of the attributes space where $f(\mathbf{X})$ returns a large positive or negative value, the minor and major radius are nearly equal (see Fig.2).

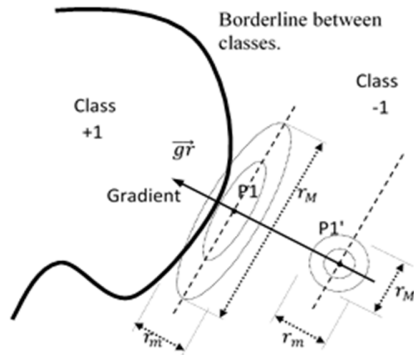


Figure 2. Equidistant distance curves (isolines) for the points P1 and P1'.

Therefore, the metric depends on just two parameters that are adjustable for each scenario: r and τ .

A. Metric properties

A differential element (line element) in a P -dimensional Euclidean space could be represented by:

$$ds = [dx_1, dx_2, \dots, dx_P] \quad (3)$$

The normalized gradient of the decision function for a certain point of the attributes space is:

$$gr = [gr_1, gr_2, \dots, gr_P] \quad (4)$$

Thus, the projection of the differential element in the direction of the gradient:

$$ds_{gr} = \langle ds, gr \rangle = \sum_{i=1}^P gr_i \cdot dx_i \quad (5)$$

And its squared norm ds_{gr}^2 in matrix form:

$$[dx_1 \ dx_2 \ \dots \ dx_P] \begin{bmatrix} gr_1^2 & gr_1 \cdot gr_2 & \dots & gr_1 \cdot gr_P \\ gr_1 \cdot gr_2 & gr_2^2 & \dots & gr_2 \cdot gr_P \\ \dots & \dots & \dots & \dots \\ gr_1 \cdot gr_P & gr_2 \cdot gr_P & \dots & gr_P^2 \end{bmatrix} \begin{bmatrix} dx_1 \\ dx_2 \\ \dots \\ dx_P \end{bmatrix} \quad (6)$$

The squared norm of the projection of the differential element on the plane perpendicular to the gradient: ds_{pg}^2

$$ds_{pg}^2 = ds^2 - ds_{gr}^2 = [dx_1 \ dx_2 \ \dots \ dx_P] \begin{bmatrix} 1 - gr_1^2 & -gr_1 \cdot gr_2 & \dots & -gr_1 \cdot gr_P \\ -gr_1 \cdot gr_2 & 1 - gr_2^2 & \dots & -gr_2 \cdot gr_P \\ \dots & \dots & \dots & \dots \\ -gr_1 \cdot gr_P & -gr_2 \cdot gr_P & \dots & 1 - gr_P^2 \end{bmatrix} \begin{bmatrix} dx_1 \\ dx_2 \\ \dots \\ dx_P \end{bmatrix} \quad (7)$$

From these results, it is possible to express a differential element in the LOM metric, given by (3), as:

$$ds^2 = [dx_1 \ dx_2 \ \dots \ dx_P] G \begin{bmatrix} dx_1 \\ dx_2 \\ \dots \\ dx_P \end{bmatrix} \quad (8)$$

$$G = \begin{bmatrix} \left(\frac{1}{r_m^2} - \frac{1}{r_M^2}\right) gr_1^2 + \frac{1}{r_M^2} & \left(\frac{1}{r_m^2} - \frac{1}{r_M^2}\right) gr_1 \cdot gr_2 & \dots & \left(\frac{1}{r_m^2} - \frac{1}{r_M^2}\right) gr_1 \cdot gr_P \\ \left(\frac{1}{r_m^2} - \frac{1}{r_M^2}\right) gr_1 \cdot gr_2 & \left(\frac{1}{r_m^2} - \frac{1}{r_M^2}\right) gr_2^2 + \frac{1}{r_M^2} & \dots & \left(\frac{1}{r_m^2} - \frac{1}{r_M^2}\right) gr_2 \cdot gr_P \\ \dots & \dots & \dots & \dots \\ \left(\frac{1}{r_m^2} - \frac{1}{r_M^2}\right) gr_1 \cdot gr_P & \left(\frac{1}{r_m^2} - \frac{1}{r_M^2}\right) gr_2 \cdot gr_P & \dots & \left(\frac{1}{r_m^2} - \frac{1}{r_M^2}\right) gr_P^2 + \frac{1}{r_M^2} \end{bmatrix} \quad (9)$$

Decomposing G into $G_1 + G_2$:

$$G_1 = \left(\frac{1}{r_m^2} - \frac{1}{r_M^2}\right) \begin{bmatrix} gr_1^2 & gr_1 \cdot gr_2 & \dots & gr_1 \cdot gr_P \\ gr_1 \cdot gr_2 & gr_2^2 & \dots & gr_2 \cdot gr_P \\ \dots & \dots & \dots & \dots \\ gr_1 \cdot gr_P & gr_2 \cdot gr_P & \dots & gr_P^2 \end{bmatrix} \quad (10)$$

$$G_2 = \frac{1}{r_M^2} I_P$$

Analyzing this G matrix, it can be concluded that G is the sum of a resultant matrix of a dyadic product: G_1 , with a scalar matrix G_2 . G_1 has only one non-zero eigenvalue equal to $\left(\frac{1}{r_m^2} - \frac{1}{r_M^2}\right) \|gr\|^2 \geq 0$. The P eigenvalues of G_2 are all equal to $\frac{1}{r_M^2} > 0$.

According to the Weyl's theorem, which states that if A and B are two symmetric matrices of dimension $P \times P$ with eigenvalues $\lambda_1(A) \leq \lambda_2(A) \leq \dots \leq \lambda_P(A)$ and $\lambda_1(B) \leq \lambda_2(B) \leq \dots \leq \lambda_P(B)$ respectively, and if the eigenvalues of the matrix resulting from the sum $A + B$ are: $\lambda_1(A + B) \leq \lambda_2(A + B) \leq \dots \leq \lambda_P(A + B)$, it is true that $\forall i \ 1 \dots P$:

$$\lambda_i(A + B) \geq \begin{cases} \lambda_i(A) + \lambda_1(B) \\ \lambda_{i-1}(A) + \lambda_2(B) \\ \dots \\ \lambda_1(A) + \lambda_i(B) \end{cases} \quad \lambda_i(A + B) \leq \begin{cases} \lambda_i(A) + \lambda_P(B) \\ \lambda_{i+1}(A) + \lambda_{P-1}(B) \\ \dots \\ \lambda_P(A) + \lambda_i(B) \end{cases}$$

From the left side inequality of the expressions above; as all the $\lambda_i(A) \geq 0$ and $\lambda_i(B) > 0$, it is concluded that all the eigenvalues of the matrix G of this metric are positive, therefore the matrix is defined positive, and its rank is P .

Alternatively to the Weyl's theorem, eq. (9) and (10) can be interpreted as the regularization of the singular matrix G_1 by means of the sum of a scalar matrix G_2 .

Spaces in \mathcal{R}^P in which the differential distance is measured using an expression of the form (8), where G is symmetric, differentiable at least twice, and its determinant is nonzero, are called Riemann spaces, and G is its fundamental or metric tensor (covariant tensor of second order).

B. Distance between two points

The length of an arc of a curve between two points is calculated by:

$$L = \int_{\lambda_i}^{\lambda_f} ds = \int_{\lambda_i}^{\lambda_f} \sqrt{\sum_{i,j} g_{ij} dx^i dx^j} \quad (11)$$

performing a parametric integration along the geodesic that connects both points.

In any metric, the distance between two points should be measured by the shortest route. A geodesic is the curve that, for two points sufficiently closed, its length is minimal among all the curves joining these two points.

From (11), and using the calculus of variations, it is shown that a geodesic should fulfil the following differential equation:

$$\ddot{x}_k + \sum_{i,j} \Gamma_{ij}^k \dot{x}_i \dot{x}_j = 0 \quad (12)$$

The derivatives are with respect to (w.r.t.) the parameter of integration and the Γ_{ij}^k are the 2nd kind Christoffel symbols:

$$\Gamma_{ij}^k = \frac{1}{2} \sum_{l=1}^n g^{kl} \left(\frac{\partial g_{jl}}{\partial x_i} + \frac{\partial g_{il}}{\partial x_j} - \frac{\partial g_{ij}}{\partial x_l} \right) \quad (13)$$

where:

- g_{ij} are the elements of the covariant metric tensor.
- g^{ij} are the elements of contravariant tensor. This tensor is computed by inverting the matrix G (which is always possible since G is nonsingular).

IV. THE LOM METRIC WHEN USING A SEPARATION FUNCTION FROM A SVM

Consider a support vector machine as the prior classification algorithm. The expression of the classification decision function is:

$$f(\mathbb{X}) = \sum_{i=1}^{nSV} \alpha_i K(\mathbb{X}, \mathbb{X}_{SV_i}) + b \quad (14)$$

where:

- nSV is the number of support vectors (SV).
- α_i is the weight of the i^{th} SV.
- \mathbb{X} is the new point to be classified.
- \mathbb{X}_{SV_i} is the i^{th} SV.
- $K(\mathbb{X}, \mathbb{X}_{SV_i}) = e^{-\frac{\|\mathbb{X} - \mathbb{X}_{SV_i}\|^2}{\sigma^2}}$ is the RBF kernel function that is applied on the \mathbb{X} point and the i^{th} SV
- b is the independent term that adjusts the decision function to be 0 in the frontier between classes.

Now, the different terms that are used in (9) are derived. The gradient (un-normalized), and its derivative w.r.t. x_k :

$$Gr_j = \frac{\partial f(\mathbb{X})}{\partial x_j} = \sum_{i=1}^{nSV} \alpha_i K(\mathbb{X}, \mathbb{X}_{SV_i}) \cdot \left(\frac{-2}{\sigma^2} \right) (\mathbb{X}_{x_j} - \mathbb{X}_{SV_{ix_j}}) \quad (15)$$

$$\frac{\partial Gr_j}{\partial x_k} = \sum_{i=1}^{nSV} \alpha_i K(\mathbb{X}, \mathbb{X}_{SV_i}) \left(\frac{-2}{\sigma^2} \right) * \left[\left(\frac{-2}{\sigma^2} \right) (\mathbb{X}_{x_j} - \mathbb{X}_{SV_{ix_j}}) (\mathbb{X}_{x_k} - \mathbb{X}_{SV_{ix_k}}) + \delta_{jk} \right] \quad (16)$$

The normalized gradient, and its derivative w.r.t. x_k :

$$gr_j = \frac{Gr_j}{\sqrt{\sum_{i=1}^P Gr_i^2}} \quad (17)$$

$$\frac{\partial gr_j}{\partial x_k} = \frac{\frac{\partial Gr_j}{\partial x_k} \sqrt{\sum_{i=1}^P Gr_i^2} - Gr_j \left(\sum_{i=1}^P Gr_i \frac{\partial Gr_i}{\partial x_k} \right)}{\sum_{i=1}^P Gr_i^2} \quad (18)$$

According with the LOM metric definition (2):

$$\frac{\partial \left(\frac{1}{r_M^2} \right)}{\partial x_j} = \frac{-2}{r_M^3} \frac{\partial r_M}{\partial x_j} \quad \frac{\partial \left(\frac{1}{r_M^2} - \frac{1}{r_M^2} \right)}{\partial x_j} = \frac{2}{r_M^3} \left(\left[\frac{r_M}{r} \right]^4 + 1 \right) \frac{\partial r_M}{\partial x_j} \quad (19)$$

$$\frac{\partial r_M}{\partial x_j} = -2 r t e^{-f^2(\mathbb{X})} f(\mathbb{X}) Gr_j = -2(r_M - r) f(\mathbb{X}) Gr_j \quad (20)$$

Also, it is possible to calculate the derivatives of the elements of the metric tensor w.r.t. the different coordinates:

$$\frac{\partial g_{ij}}{\partial x_k} = \frac{2}{r_M^3} \left(\left[\left(\frac{r_M}{r} \right)^4 + 1 \right] g_{ri} g_{rj} - \delta_{ij} \right) \frac{\partial r_M}{\partial x_k} + \left(\frac{1}{r_M^2} - \frac{1}{r_M^2} \right) \left(g_{ri} \frac{\partial g_{rj}}{\partial x_k} + g_{rj} \frac{\partial g_{ri}}{\partial x_k} \right) \quad (21)$$

With all the above results, it is possible to calculate the Christoffel symbols and the geodesic that connects two points.

V. DISTANCE BETWEEN POINTS IN THE LOM METRIC

There are two major approaches to calculate distances based on a metric that varies locally: to approximate the integration of (12), or employ an algorithm that evaluates distances for a grid of points and then, choose the shortest path between origin and destination.

A. Distance calculated by integrating the geodesic

The integration of (12) can be accomplished by many methods. Perhaps the simplest one is:

- Discretize the path between points in a finite number of points: x^m , $0 \leq m \leq M$. The initial and final points x^0 and x^M , are fixed points. x_k^m is the k^{th} coordinate of the x^m point.
- Set the equivalent differences equation at each point:

$$x_k^m \leftarrow \frac{(x_k^{m+1} + x_k^{m-1})}{2} + \frac{1}{8} \sum_{i,j} \Gamma_{ij}^k(x^m) (x_i^{m+1} - x_i^{m-1})(x_j^{m+1} - x_j^{m-1}) \quad (22)$$

- Iterate until the equation is satisfied for all the points within a preestablished error.
- The resulting points profile the geodesic between the initial and final points.

B. Distance calculated by the shortest path

To integrate the differential equation of the geodesic does not guarantee finding the shortest path between two points. A solution that guarantees it, is to search for the shortest path by a well-known algorithm as the one proposed by Dijkstra [17].

VI. MEASURES OF PERFORMANCE

A two attributes synthetic problem is defined. It resembles an ill-behaved, non-linearly separable classification problem, with sufficient complexity to be interesting. It could be easily reproducible and the number of cases may vary at will in a repeatable way. Thus, a two class problem with six hundred cases for each class is generated:

- The first class consists of data taken from three independent Gaussian distributions with means:

$$\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \begin{bmatrix} 4 \\ 4 \end{bmatrix}$$

and covariance matrices:

$$\begin{bmatrix} 0.1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 0.1 \end{bmatrix}, \begin{bmatrix} 1 & -\sqrt{1/2} \\ -\sqrt{1/2} & 1 \end{bmatrix}$$

- The second class is formed with an equal amount of data, but taken from only one Gaussian distribution with mean and covariance: $\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

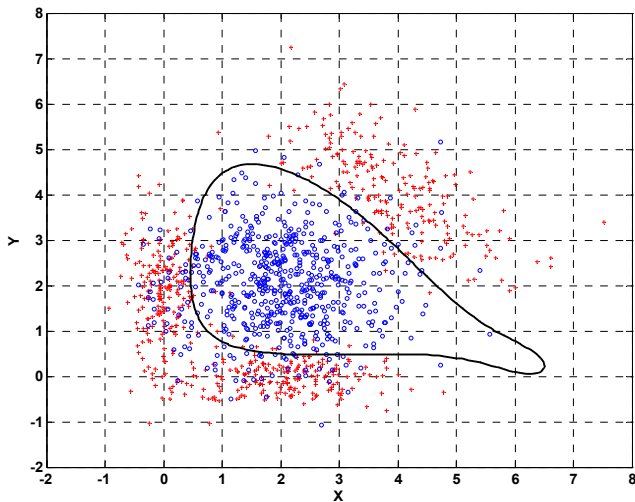


Figure 3. Graphical representation of the synthetic problem. The round points belong to the class -1, the crosses to the class +1.

Fig. 3 provides a graphical representation of these six hundred cases. It also includes a closed curve that depicts the class separation function obtained from a RBF-SVM (with parameters $C=4, \gamma=2$).

One of the advantages of a synthetic problem is that it is possible to calculate the expected error for classifying a new case. In our problem, the Bayes error is 10.22%.

By a ten-fold cross-validation technique, the parameters of a SVM based on a RBF kernel are adjusted. The optimal parameters for this SVM are: $C = 4, \gamma = 2$, showing an average error in the classification of 10.75%. Therefore, it is only at 0.53% of what can be considered the optimal classification.

A. Geodesic curves in the LOM metric for this test bed

Using the decision curve provided by the SVM, the LOM metric can be defined. Then, it is possible to calculate the distances from an arbitrary point (in Fig. 4 is the point [0,4]) to any other points in the plane by integrating the differential equation of the geodesic (22).

It can be clearly seen, that the traditional straight lines between points in a Euclidean metric have been replaced by curves that follow profiles similar to the decision function to reach their destinations.

This means, that the shorter paths among points prefer to surround the surface of separation between classes (instead of crossing it). Therefore, under this metric, the points that are located in paths parallel to the boundary show greater similarity, and smaller distance, to the case to be classified.

The three main drawbacks of this technique are:

- A geodesic trajectory is not guaranteed to be the shortest path.
- The time required for calculating the distance among all the cases of the problem. A differential equation must be integrated for each pair of points.
- The geodesic trajectory may vary depending on the initial path considered for the integration of the differential equation.

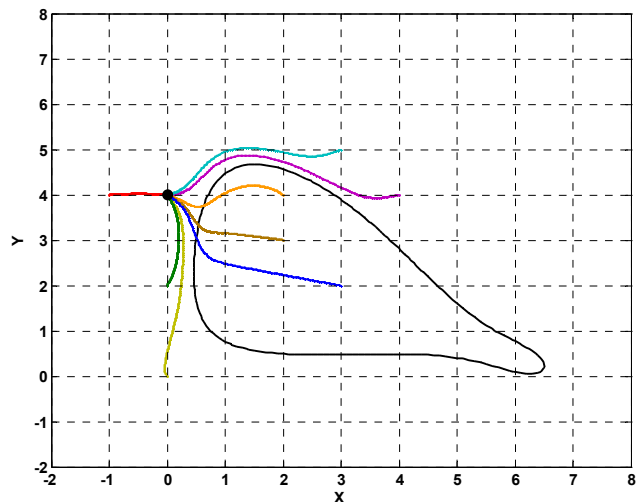


Figure 4. Geodesic curves to reach different points from (0,4) according with the LOM metric.

B. Dijkstra's shortest path between two points

In this research, we have implemented a variation of the Dijkstra's algorithm that uses a priority queue to accelerate calculations and allows diagonal paths between the elements of the grid. The results can be seen in Fig. 5. The results agree with those obtained by integrating the geodesic equation.

In this case, it is possible to calculate the distance between one point and the rest of the points in the space of attributes; and it is also easy to draw the contour of the isolines for a given point.

Thus, the objective has been achieved, the isolines of the metric are not simple circles, nor ellipsoids fixed at the origin (as it happens in the current most advanced algorithms), but curves better adapted to each problem. It can be seen that trips to relatively close or distant points according to the Euclidean metric, become larger or smaller distances according in which direction is travelled from the starting point.

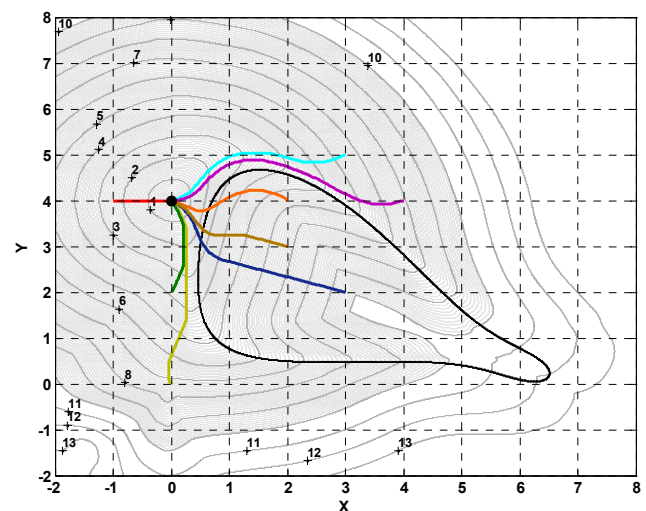


Figure 5. Shortest paths to reach different points from (0,4) using the LOM metric. It can be seen also the isolines that join equidistant points.

C. Using the LOM metric in the k-NN algorithm

We prepared four sets of 180, 90, 60 and 30 cases of the two class problem described in Section VI.A to train four different k-NN classifiers to test the performance of the new metric; and another test set (independent of those enumerated above) containing over 6000 cases. The aim was to show not only that the classification accuracy improves when using the LOM metric, but also that this is true when either the number of training cases is high or low.

Therefore, different SVMs for each problem were trained, and then the k-NN algorithm was applied (using both the

Euclidean and LOM metrics) to classify each of the 6,000 test cases. In this study, to calculate the distance between two points with the LOM metric, the Dijkstra technique was used (we also employed different edge lengths to study the effect of degeneration of distance measurements when this parameter got larger). In Table 1, and in the four graphs of Fig. 6, the results obtained are shown.

In each of the four scenarios (180, 90, 60 and 30 cases), a significant increase of classification accuracy is achieved with the k-NN algorithm which employs the new LOM metric (compared with the results obtained with the k-NN or RBF-SVM algorithms using the Euclidean metric).

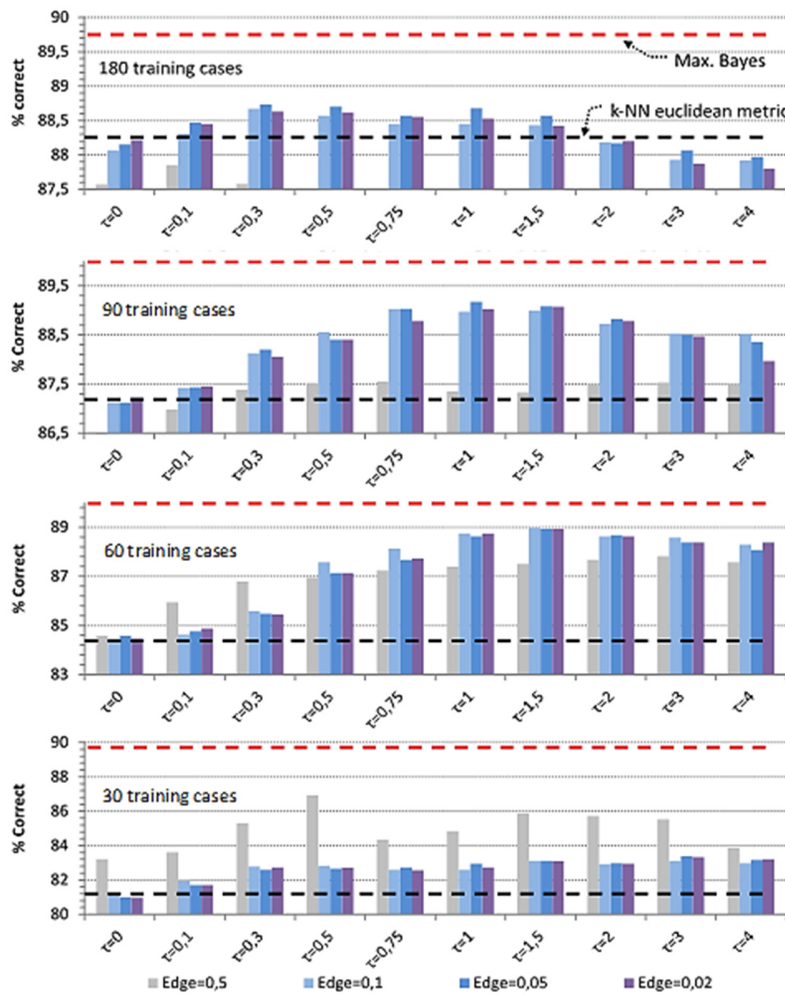


Figure 6. Results of the k-NN classification for the four problems.

TABLE I. SUMMARY OF CLASSIFICATION ACCURACY

Problem	k-NN euclidean metric	SVM	k-NN LOM metric
180 cases	88.25% (k=13)	88.54%	88.65% ($\tau=0.2, k=13$)
90 cases	87.23% (k=3)	87.53%	89.12% ($\tau=1.25, k=7$)
60 cases	84.40% (k=5)	85.08%	88.98% ($\tau=1.25, k=13$)
30 cases	81.23% (k=1)	82.40%	83.33% ($\tau=3, k=2$)

VII. CONCLUSION AND FUTURE WORK

The main objective of the LOM metric has been achieved; it does not only enlarge the space parallel to the decision function, but it also provides a metric whose isolines are not ellipses centered at the initial point that extend along all the space, but curves which adapt locally to the profile of the decision function.

As it is frequently remarked in the literature, the shortest path between points in a Riemannian metric is a geodesic, but not every geodesic is the shortest path. Several times in this research, we have found two different geodesics to reach the same point. This is due to the different initial path used, and the attractors that govern the integrations of the differential equations. Under this point of view, it is preferable to use the Dijkstra's algorithm.

However, one of the main drawbacks of Dijkstra's algorithm is the time required to calculate the distance between points. Its order of complexity is $O((|E| + |V|)\log|V|)$ using a priority queue, being $|E|$ the cardinality of edges and $|V|$ the same measure for the vertices.

Analyzing the results obtained with the k-NN algorithm, some aspects become relevant:

- It is not necessary to explore different values for the r parameter (for k-NN, r is just a scale factor).
- The range of the τ parameter for which significant improvements are obtained is wide. This is because the LOM metric is derived from a well-founded theoretical concept.
- Even for edges of 0.1 units, the distance calculations by the Dijkstra's algorithm are correct enough (the range of the attributes in each dimension is between -1 and 6). Shorter edges do not provide a significant improvement of accuracy and only increase calculation time.

One aspect for further research is the study of the degradation of Dijkstra's algorithm when used in spaces with more dimensions than two. But, perhaps the most relevant area of improvement is to define a simplification of the LOM metric that allows its use in multidimensional spaces with reasonable time and memory constraints. We are currently working in one of them and our first results are very promising.

As a conclusion, calculation of distances using the LOM metric is seen as an alternative for those algorithms that use this measurement in their decision making. It increases the classification accuracy and reduces the curse of dimensionality. The main drawback of the LOM metric is that more extensive calculations are needed to obtain the distances. Currently, we are working to minimize this inconvenience.

REFERENCES

- [1] R. Duda, P. Hart, and D. Stork, "Pattern Classification" 2nd Ed. John Wiley, New York, 2000.
- [2] C. Domeniconi, D. Gunopulos, and J. Peng, "Large Margin Nearest Neighbor Classifiers", IEEE transactions on Neural Networks, vol.16, 2005, pp. 899-909.
- [3] J. Peng, D. R. Heisterkamp, and H. K. Dai, "LDA/SVM Driven Nearest Neighbor Classification", IEEE Transactions on Neural Networks, vol.14, 2003, pp. 940-942.
- [4] J. Revilla and E. Kahoraho, "BTW: a New Distance Metric for Classification", Proc. of the International Symposium on Distributed Computing and Artificial Intelligence, DCAI 2012, 2012, pp. 701-708.
- [5] K. Q. Weinberger and L. K. Saul, "Distance Metric Learning for Large Margin Nearest Neighbor Classification", Journal of Machine Learning Research, Vol.10, 2009, pp. 207-244.
- [6] S. Shalev-Shwartz, Y. Singer, and A. Y. Ng, "Online and Batch Learning of Pseudo-metrics". Proc. of the 21st Int. conference on machine learning, Banff, Canada, 2004, pp. 94-101.
- [7] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. "Learning a Mahalanobis metric from equivalence constraints". Journal of Machine Learning Research 6, 2006, pp. 937-965.
- [8] N. Shental, T. Herz, D. Weinshall, and M. Pavel, "Adjustment learning and relevant component analysis", Proc. of the 7th European conference on computer vision, London, UK, 2002, pp. 776-790.
- [9] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood Component Analysis", Advances in neural information processing systems vol. 17, 2005, pp. 13-520.
- [10] Q. Chang, Q. Chen, and X. Wang, "Scaling Gaussian RBF Kernel Width to Improve SVM Classification", Int. Conf. on neural networks and brain, ICNN&B '05 vol. 1, 2005, pp. 19-22.
- [11] S. Amari and S. Wu, "Improving Support Vector Machine Classifiers by Modifying Kernel Functions", Neural Networks vol.12, 1999, pp. 783-789.
- [12] S. Wu and S. Amari, "Conformal Transformation of Kernel Functions: a Data-Dependent Way to Improve Support Vector Machine Classifiers", Neural processing letters, vol. 15, 2002, pp. 59-67.
- [13] G. Wu and E. Chang, "Adaptive Feature-space Conformal Transformation for Imbalances-data Learning", 20th Int. Conf. on Machine Learning (ICML-2003), 2003, pp. 816-823.
- [14] P. Williams, S. Li, J. Feng, and S. Wu, "Scaling the Kernel Function to Improve Performance of the Support Vector Machine", Advances in Neural Networks, ISNN'05, 2005, pp. 831-836.
- [15] F. Fernandez and P. Isasi, "Local Feature Weighting in Nearest Prototype Classification", IEEE Transactions on Neural Networks, vol.19, 2008, pp. 40-53.
- [16] Y. Zhang, H. Zhang, N. M. Nasrabadi, and T. S. Huang, "Multi-metric Learning for Multi-sensor Fusion Based Classification", Information Fusion, vol. 14, 2013, pp. 431-440.
- [17] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, "Introduction to Algorithms", MIT Press, Massachusetts. 2000.

First Steps Towards a Formal Analysis of Law

Tom Maarten van Engers
Leibniz Center for Law
University of Amsterdam
Amsterdam, the Netherlands
e-mail: vanengers@uva.nl

Robert van Doesburg
Immigration and Naturalization Service
Rijswijk, the Netherlands
e-mail: r.v.doesburg@ind.minvenj.nl

Abstract—In this paper, the authors present some work recently done within the Dutch Immigration and Naturalization Service (IND). Being responsible for the implementation and execution of complex and ever changing regulations, for which the use of IT systems is a necessity, this organization has become aware of their dependence on trustworthy methods to assure the correct implementation of law into their operations and services. While many attempts to automate law, even in the domain of migration law, have been made before, hardly any attention has been paid to the ‘translation process’ from legal rules expressed in natural language to specifications in computer executable form. In this paper, we will explain the method we have developed and illustrate its application with some concrete examples. The work is part of a larger innovation programme initiative that we collaboratively conduct within a virtual collaboration, called the ‘Blue Chamber’.

Keywords—knowledge acquisition; legal engineering; legal analysis; Hoffeldian analysis.

I. INTRODUCTION

Making a formal analysis of law is problematic because sources of law contain huge amounts of implicit information. As a result, law can be difficult to understand and to interpret. In countries that follow common law tradition understanding the law requires the knowledge of a growing body of potentially relevant preceding cases, next to the knowledge of ‘black letter’ law, including bills such as tax law, or immigration law. The authors, living in a civil law country, conduct their research in governmental organizations. In civil law cultures, the most important legal sources are the laws that are produced by the parliamentary system and the sources of law that are produced by legal bodies based on delegated powers. Case law is relevant as well, especially in explaining the ‘correct’ interpretation and application of legal sources onto actual cases. But rather than taking case decisions as a primary source, institutions like the IND describe the legal consequences of court decisions in changed regulation.

Translating sources of law into formal specifications for IT systems is a necessary step for (partial) automation of public services. The translation process of sources of law should be transparent to make sure that legal and policy experts are able to validate the correctness and completeness of the ‘legal rules’ the IT system. The authors have also

pointed at the importance of inter-coder independency and scalability with respect to human resources [1]. The method to be developed should allow us to work with many knowledge analysts and knowledge engineers that have normal professional skills. Like us, other people have identified the need for a systematic approach that would allow to cope with law, and enable its translation into formal models that consequently could be executed by IT systems. In fact the very idea of translating law into computational models that we could use for solving cases using a computer, goes back to Gottlieb Wilhelm Leibniz who even build a mechanical reckoner that was supposed to solve complex (legal) problems once we would have the magical numbers of the legal concepts. Leibniz before Wilkins, both deserving recognition as founding fathers of modern computer science, stressed the importance of formal conceptualizations, these days usually referred to as ontologies, centuries before computer science came into existence [9][17].

Shortly after the invention of the modern computer, scientist recognized its power as a symbol processor allowing it to be used for reasoning processes. This includes reasoning about legal cases. Consequently, people have been working on using IT systems within the domain of law, particularly within public administrations. In literature, one can find quite some research papers on systems aimed at deciding on legal cases. The authors of this paper also have been working on such legal decision-support systems [2][3].

The early work of Sergot et al., which is also within the domain of immigration law, like the work we present in this paper, uses the expert as the main source of legal knowledge, despite of its title: ‘the British Nationality Act as a Logic Program’ [13][14]. In the POWER-programme [4], conducted in the late nineteen-nineties by the authors, written sources of law were the main source of (legal) knowledge. Legal experts did, and do still, play an important role, but only as interpreters and validators of knowledge that can be traced back to sources of law.

In a series of projects since the POWER-programme, the authors have been systemizing the translation process using (semi) automated norm extraction [1][11][12], particularly by looking at invariant language patterns typically used in written sources of law. Using computational linguistics we were able to identify the most important patterns and showed that we could use parsing to ‘translate’ written sources of law

written in natural language into model sentences in a formal language.

Also, we have been working on the representation of norms in ways that enable multiple task contexts and multiple agents perspectives. The typical single task orientation used in Sergot et al. could be avoided. In order to achieve this objective, we used formal models [5][6] that are based upon an extended version of Hohfeld's model [8]. In previous research projects we have showed that Hohfeld's initial model is already a big improvement compared to traditional interpretation of rights, duties and allowances in (modal) logic. Our extended and completely formalized version of Hohfeld's model enables us to express all typical jurial relations in a formal way, but it is also expressed in a relational model that can be implemented in a straight forward way. Furthermore, we worked on the development of an agent-role based model, allowing us to reason about the consequences of norms in a social context [15][16]. We certainly do not claim that we have solved all issues, but by working on these related topics the depth and complexity of understanding law became much clearer to us and has inspired us to continue our quest.

One of the issues that we did not address thus far was the scoping problem. While legal experts are perfectly able to list the regulations that are relevant to solving a legal problem of some kind (or at least claim to be), we have experienced that one of the problems for our knowledge analysts was where to start analyzing, what rules of law to include, and where to stop looking for additional sources of law. Obviously, a top down approach, analyzing all sources of law in a country, could not possibly work since there are simply too many sources of law available to allow for an analysis that finishes within a reasonable time. Furthermore, the concept 'top down' would be problematic. Although the constitution normally is considered to be the highest source of law (in the Netherlands just after 'the grace of God' that gives the formal power to the King), we are bound by even higher forces, such as international treaties. In the POWER-programme we have experienced that a serial approach, where analysts worked through a source of law from the first to the last article, was both time consuming and required an integration step to 'glue' the different partial models together. At that time we did not have an explicit method for doing that and much was left to the insights of the analysts.

So the challenge we took up was to make explicit the issue of scoping relevant sources of law. If we would have a method that solves our scoping problem in such way that it is coder independent, and results in a model that could be mapped to the original sources (like we aimed for in the POWER-programme), we would be a step closer to our final aims, i.e. a method for the formal analysis of law.

In this paper, we describe the approach we have developed and explain it by illustrating its application on Dutch Alien Law. We will explain the issues that were raised during the analysis and explain their relevance.

In Section 2, we give a short description of recent IT developments within the IND. In Section 3, we give a brief overview of legal sources relevant to the domain. In Section 4, we present a scoping procedure for sources of law. Section

5, briefly describes the conceptual-semantic analyses. Section 6, contains conclusion and future work.

II. INNOVATION OF THE IND

In 2005 the IND has chosen to fundamentally rebuild its organization. One of the steps toward a newborn IND was the redesign of all processes, including all supporting IT systems. This new system, called INDiGO, would be based on rule governance principles, separating procedural knowledge (workflow) from legal case content related knowledge. This principle became known as separating the know from the flow [7]. In order to create the actual IT solution, the IND selected several middle ware components, following the architectural principle of having one type of functionality within one middle ware component (one thing in a box). In order to support its knowledge management the IND has chosen to work with an inference engine that works on an explicit knowledge model derived from sources of law. This set up was also intended to enhance adaptivity and reduce maintenance efforts as a result of changing law and policies. Frequent changes of the knowledge model should not compromise the systems stability.

The INDiGO system is now operational for more than five years, and while the maintenance efforts are significantly less compared to the previous IT-systems, the effort required to implement changes is still substantially higher than expected when the solution was chosen.

The biggest issue is that legal experts and policy advisors lack the skills or will to read and validate the knowledge representation used by the inference engine in order to adapt changes when necessary. This is not a new problem. The Common Business Oriented Language (COBOL) was also once expected to be used by business people, rather than the very specialized computer programmers that actually created the systems code.

If systems are relatively stable the absence of direct insight of legal experts and policy advisors in the knowledge models would not hurt too much. But in volatile environments with constantly changing laws and policies, organizations will loose control of their IT systems unless they find ways to support the 'translation' of legal knowledge from the sources of law into a formal representation in their IT systems.

As a result of not using the knowledge models in the inference engine the user organization of the IND specifies changes without direct insight into the existing set of rules. Integrating the amended rules in the existing knowledge models is left to knowledge modelers. Although knowledge modellers within the IND all have a history in the primary process of the organization, this leads to an unnecessary burden on the IT change process. Because requests for change in the inference engine are not fully specified. Interpretation errors made by knowledge workers, because of the lack of details in specifications, will not be detected until the start of the test phase. This leads to relatively expensive modifications.

The absence of unambiguous specifications also has consequences for the acceptance of changes by users. Having formal specifications is a condition for making a complete

test set. Test employees lacking full understanding of the required specifications will by definition carry out incomplete acceptance tests. Recognizing errors in the last stages of a release, or even in the initial period after delivery, leads to even higher repair cost than corrections made during the design and testing phase of a release.

The IND is seeking a formal method for coder independent and traceable specifications for the implementations of policy changes and changes in sources of law in information systems. These specifications should be available at the start of the process of changing knowledge models.

The authors have not found any existing solution for solving this issue.

III. LAWS AND REGULATIONS IN THE NETHERLANDS AND EUROPE

Why is scoping difficult? Laws are referring, explicitly and implicitly, to other laws and subordinate legislation. This creates a network of relationships. Determining relevant relationships can only be done on the basis of a given context. The scoping process should lead to a set of rules that can be traced back to sources of law and to a context description.

On October 15, 2014 the following rules were effective in the Netherlands, and thus potentially relevant for the IND:

- 1.100 Dutch Laws;
- 1.748 Dutch Orders in Counsel
- 5.273 Dutch ministerial regulations
- 679 Dutch international treaties
- 2.796 European Laws
- 3.164 European regulations
- Dutch and European case law

The IND is responsible for the implementation of the Aliens Act and the Netherlands Nationality Act. This article focuses on the implementation of the Aliens Act. The Aliens Act consists of 173 articles, 288 lines, 28.766 words. Subordinate legislation includes:

- Order in Council: the Aliens Decree (322 articles, 5.642 lines, 56.331 words)
- Ministerial regulation: Aliens Regulations (167 articles, 3.983 lines, 27.213 words)
- Aliens Act Implementation Guidelines, part A to D (7.491 paragraphs, 13.935 lines, 153.068 words).

The challenge of the scoping process is to distil a relevant set of rules out of the entire collection of sources of law.

IV. A PROCEDURE FOR SCOPING

To select a workable set of rules that provides an adequate basis for developing specifications for a specified context, we used the following procedure:

1. Select an acting person (this can also be an organization). Describe the context of the acts of the agent
2. Choose a starting point for the analysis: a legal statement that contains a condition and is relevant in the chosen context.
3. Perform a linguistic analysis on a selected article.

4. Transform the text to the active voice, thus insuring there the subject of the sentence is an acting person.
5. Identify explicit references and terms that need a definition.
6. Select words or constituents that contain or might contain an implicit reference. Make these references explicit, or make an explicit decision that further analysis is not relevant in the chosen context.
7. Analyze all the selected words and constituents, starting with point 2 of this procedure.
8. The procedure ends when all relevant references analyzed. The decision to end the analysis is being made by a multidisciplinary team in which legal experts, policy advisors, and practitioners are represented.

The reader must be aware that the linguistic analyses (step 3 of the procedure) for the example cases described below have been conducted manually. The authors aim at supporting this step by automated devices in the near future and have worked on automated tools for the analysis of legal sources in the various previous projects [10][11]. The examples described below show that even a quite limited lingual analyses is sufficient for our purpose.

A. Case: foreign students in the Netherlands

The procedure for scoping has been tested for the context of services provided by the IND (acting person) to foreign students studying or wanting to study in the Netherlands at a university.

B. Starting the analysis

Choosing a starting point for the analysis is step 2 of the scoping procedure. Two logical starting points for the analysis are:

1. A foreign student wants to come to the Netherlands. Conditions for admission to the Netherlands are stated in article 3, Aliens Act (Chapter 3: Entry).
2. A foreign student wants to reside in the Netherlands. Conditions for residence in the Netherlands are stated in article 8, Aliens Act (Chapter 3: Residence).

Both starting points lead to the same results within five iterations, because of implicit cross-references between article 3 and article 8, Aliens Act.

C. The analysis of an article

Step 3 of the procedure, the linguistic analysis of a selected article, is illustrated on the basis of Aliens Act article 16, paragraph 1, preamble and under b and article 4: 1 of the General Administrative Law.

Aliens Act article 16, paragraph 1, preamble and under b reads: "An application for a temporary residence permit can be rejected, if the alien does not possess a travel document." This is a sentence in the passive form and it contains the following components, see Figure 1:

- can be rejected (verb)
- an application for the granting of a residence permit for a fixed period as referred to in article 14 (subject)
- if (conditional conjunction)

- the alien does not possess a travel document (adverbial of condition; subordinate clause).

The subordinate clause of the sentence can be decomposed to:

- possess (verb)
- not (adverb of denial)
- the alien (subject)
- a valid travel document (direct object).

Because the sentence is in the passive voice there is no acting subject present.

In step 4 of the procedure the sentence is put in the active voice. Resulting in the question: who is the acting person that can reject an application for a residence permit.

The question is answered in step 5 of the procedure. The explicit reference to Aliens Act article 14, paragraph 1 reveals that it is Our Minister (the minister of Security and Justice) who is authorized to grant, reject or disregard an application for a residence permit.

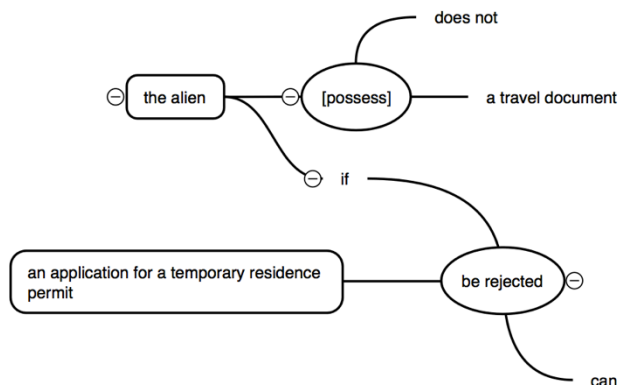


Figure 1. Linguistic analysis of Aliens Act article 16, paragraph 1, preamble and under b.

After annotating the explicit references in the article, step 6 of the procedure commences: the search for implicit references. There is an implicit relationship between ‘an application for a temporary residence permit’ and ‘the alien’ that has to possess a travel document. Also the existing of ‘an application for a temporary residence permit’ implies that such an application can be submitted. To be able to make this implicit reference explicit, relevant rules are searched for in sources of law.

To make explicit what ‘the alien’ should do to proof he possesses a valid travel document, the requirements that enable ‘Our Minister’ to assess whether ‘the alien’ has a valid travel document are added, including references to the relevant sources of law, see Figure 2.

Step 7 of the procedure is the analyses of the articles of sources of law to which implicit or explicit references have been found.

The procedure continues with the analysis of article 4:1 of the General Administrative Law is shown, see Figure 3.

Article 4:1 of the General Administrative Law reads: “The application for taking a decision is submitted in writing to the administrative authority authorized to decide on the application, unless otherwise prescribed by law.” The sentence is in the passive form and contains the following components:

- is submitted (verb)
- the application for taking a decision (subject)
- unless (conditional conjunction)
- otherwise prescribed by law (adverbial of condition)
- to the administrative authority authorized to decide on the application (adverbial of place)
- in writing (adverbial of condition)

Repeating the scoping procedure leads to new implicit references, see Figure 4.

D. The size of the set of rules

The scoping procedure leaves 38 articles relevant for the entry and residence of foreign students in the Netherlands

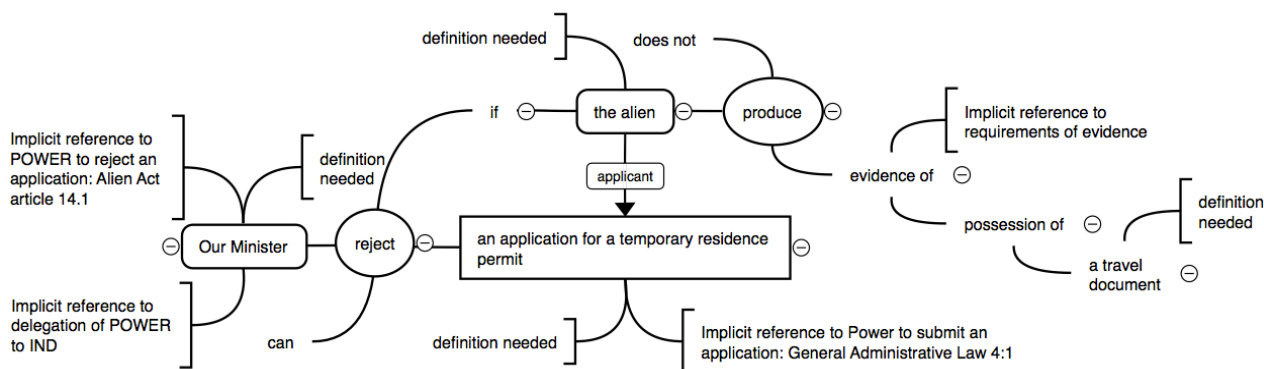


Figure 2. Transformation of Aliens Act article 16, paragraph 1, preamble and under b.

out of the original 173 articles of the Aliens Act. The set of rules also contains related sources of law, including rules from other legal domains.

The overall set of rules thus covers the following components (articles and paragraphs) of the regulations:

- 38 of the 173 articles of the Aliens Act
- 26 of the 322 in the Aliens Decree articles
- 12 of the 167 articles in the Aliens Regulations
- 280 of the 7.491 paragraphs of the Aliens Act Implementation Guidelines
- 1 article from a European law
- 13 articles of the General Administrative Law Act
- 2 articles from the Public Health Act
- 1 article of the Criminal Code
- 1 article from the Student Finance Act.

E. The value of a linguistic analysis

New in the proposed method is the explicit annotation of every legal rule that is needed for capturing the formal specifications of a service. Reasons for adding a legal rule to the relevant set of rules are recorded. Annotations for not adding a rule because of lack of relevance is only recorded when the analysts involved decide that this information might be relevant in the future. Every rule and every annotation in the model is explicitly related to a source of law.

The set of legal rules and the annotations made in establishing the relevant set of rules provides a basis for discussion and further analysis of the accuracy of the determined scope.

Adjusting the set of rules is possible at any time. Reapplying the procedure is necessary when changes are being made onto sources of law referred to in the relevant set of rules, or when any expert involved asks for adjustments.

The obtained legal rules must be converted to a formal (executable) specification. We use a specific application of CogNIAM (see [5][6]). This results in a formal model expressed in some relational algebra. It is therefore quite different from other specification languages that have a rule-like syntax, such as Prolog, most production rule syntaxes, business rules, etc. CogNIAM specifications can be used as formal specifications for IT systems and also provides a full set of test scenarios. With respect to the analytical approach used so far for the making of CogNIAM representations, the scoping procedure presented in this paper, yields the great

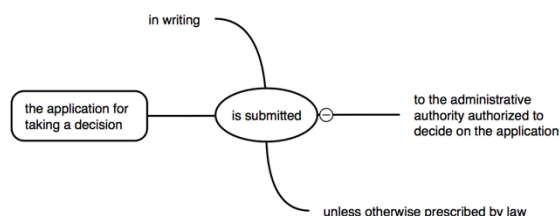


Figure 3. Linguistic analysis of article 4:1 General Administrative Law.

advantage that the linguistic analysis is easy to learn. The constituents resulting from the linguistic analysis are much more structured than the original legal text, leading to a decrease of coder independency. Additionally the full traceability to legal text enables standardization of the translation of legal text in natural language to formal language. In the near future unambiguous automated methods for linguistic analysis on legal text will be available (see also [1][11][12]). For multiple languages including Dutch and English parsers that can perform this analysis automatically are available.

The intermediate step of linguistic analysis simplifies the often complex sentence structures in legal texts and thus provides for legal experts better understanding of the intended meaning of legislature, and makes anomalies that would otherwise remain hidden, visible.

V. CONCEPTUAL-SEMANTIC ANALYSIS

In [5][6] Van Engers and Nijssen describe the conceptual analysis of law. In order to allow the reader to understand how the models that are produced follow the steps described in the previous sections, we will briefly give the Hohfeldian analysis of one of the partial models (Aliens Act article 16, paragraph 1, preamble and under b). This model (see Figure 2) describes potential legal relations between ‘Our Minister’ and ‘the alien’. In Hohfeldian terms this model is read as follows: ‘Our Minister’ has a POWER to reject a request for a temporary residence permit, when ‘the alien’ cannot produce evidence of being in possession of a travel document.

When ‘Our Minister’ executes that POWER ‘the alien’ has a LIABILITY towards the rejection and a new jural relation between ‘Our Minister’ and ‘the alien’ comes into existence, namely the DUTY of ‘Our Minister’ to produce a decision (rejection) and a (CLAIM)RIGHT of ‘the alien’ on that decision. Once ‘Our Minister’ has fulfilled his DUTY the DUTY-(CLAIM)RIGHT relation terminates.

However, if ‘the alien’ can produce evidence of being in possession of a travel document, then ‘Our Minister’ has a DISABILITY and ‘the alien’ an IMMUNITY with respect to the rejection of the request for a temporary residence permit.

The model does not answer the question what happens then? We may not infer that by not having the POWER to reject the request will then be granted. So we have to look

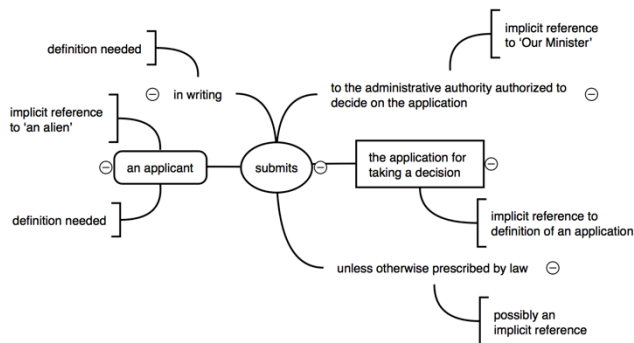


Figure 4. Translation of article 4:1 General Administrative Law.

for other partial models that will provide us an answer to this question.

Also one may wonder how 'Our Minister' would know about the alien's request unless it is submitted (by 'the alien'?). Looking for an answer to this question we might find that this is arranged for in AWB 4.1 (see the corresponding model in Figure 3). According to this model 'the alien' got a POWER to submit a request for a temporary residence permit. If 'the alien' would execute this POWER 'Our Minister' would have the LIABILITY to decide upon the request thus creating a new jural relation between 'Our Minister' and 'the alien'. This DUTY-(CLAIM)RIGHT relation would give a DUTY to 'Our Minister' to make a decision and a (CLAIM)RIGHT on that decision for 'the alien'. This jural relation terminates once 'Our Minister' makes the decision.

If we would combine all these partial model we would find that the intended scenario would be as follows: When 'the alien' executes his POWER to submit a request for a temporary residence permit, 'Our Minister' has the LIABILITY to decide upon the request. This is the first POWER-LIABILITY relation. This creates a new legal relation between 'the alien' and 'Our Minister'. Following from this relation 'Our Minister' got the DUTY to decide and 'the alien' the (CLAIM)RIGHT on the decision. When 'Our Minister' fulfills his DUTY the relation between 'the alien' and 'Our Minister' is terminated. The second POWER-LIABILITY relation is relevant for content of the decision. According to the combined model 'Our Minister' has POWER and 'the alien' has a LIABILITY towards a rejection if 'the alien' cannot produce evidence of being in possession of a travel document. 'Our Minister' then has the DUTY to produce a decision (rejection) and 'the alien' has a (CLAIM)RIGHT on that decision. Once 'Our Minister' has fulfilled his DUTY the DUTY-(CLAIM)RIGHT relation terminates. If 'the alien' produces evidence of being in possession of a travel document, then 'Our Minister' has a DISABILITY and 'the alien' has an IMMUNITY with respect to the rejection of the request for a temporary residence permit. What should happen in this case remains unclear.

The Hohfeldian analysis given above shows that the deeper analysis of the partial models further clarifies the meaning of sources of law in terms of (potential) legal consequences of agents' acts. Furthermore the analysis helps to identify issues that need further clarification to fully cover the potential situations that may occur within a given context.

VI. CONCLUSION AND FUTURE WORK

The title of this paper was deliberately chosen. One could argue that the first steps towards formal analysis of law have been set long ago. For centuries researchers have been trying to formalize law, and certainly the last decades a lot of progress has been achieved. Legal information systems have become essential components in our modern society and without them efficient and effective application of law would be impossible. Despite all of this a complete method that translates law written in natural language into formal

specifications is still lacking, although parts of such a method are already available.

The work we report on in this paper, fits within a series of studies, experiments and trials that are all aimed at developing a method for the formal analysis of law. In this paper, we addressed an issue - scoping - that has been neglected so far, but is an essential part of the method we are working on. In this sense we address the first steps in the translation towards a formal model of law.

By developing our approach and testing it on concrete cases we have learned a lot about how humans, including knowledge engineers and legal experts read and interpret sources of law and attribute meaning to them.

Following the scoping procedure explained in this paper, we have not only been able to identify relevant pieces of law, and thus helped to find a solution to our scoping issue, but while building the partial models and transforming them, we have been able to identify and clarify implicit relations between constituting parts of sources of law, terminological and conceptual unclarity and vagueness. This makes that our approach offers much more than merely filling the scoping gap in our method, as it contributes to improving the quality of the analysis of law and consequently helps public administrations such as the IND to improve the quality of their operation.

ACKNOWLEDGMENT

The research reported upon in this paper, wouldn't have been possible without the support of the Dutch Immigration and Naturalization Service. Also the cooperation in the Blue Chamber has inspired us to continue to work on this intriguing topic and discuss our ideas even in late hours and at the cost of not paying attention to our spouses.

REFERENCES

- [1] T. M. van Engers, "Legal engineering: A structural approach to improving legal quality" in A. Macintosh, R. Ellis and T. Allen, editors, *Applications and Innovations in Intelligent Systems XIII*, proceedings of AI-2005, Springer, Dec. 2005, pp. 3-10, ISBN 978-1-84628-224-9.
- [2] T. M. van Engers and E. Glassée, "Facilitating the legislation process using a shared conceptual model," in *IEEE Intelligent Systems*, 2001 vol.16 Issue No.01, pp. 50-58, ISSN: 1541-1672.
- [3] T. M. van Engers, P. J .M. Kordelaar, J. den Hartog and E. Glassée, "POWER: Programme for an Ontology based Working Environment for modeling and use of Regulations and legislation," in Tjoa, Wagner and Al-Zobaidie, editors, *Proceedings of the 11th workshop on Databases and Expert Systems Applications (IEEE)*, Greenwich London, 2000,, pp. 327-334, ISBN 0-7695-0680-1.
- [4] T. M. van Engers and P. J. M. Kordelaar, "POWER: Programme for an Ontology based Working Environment for modeling and use of Regulations and legislation," *Proceedings of the ISMICK '99*, ISBN 2-913-923-02-X.
- [5] T. M. van Engers and S. Nijssen, 2014, "From Legislation towards the Provision of Services - An Approach to Agile Implementation of Legislation," in A. Kő and E. Francesconi, editors, *proceedings of the Third International Conference on Electronic Government and the Information Systems Perspective (EGOVIS 2014)*, Springer, München, Germany, Sep. 2014 pp. 163-172, ISBN: 978-3-319-10177-4, e-ISBN: 978-3-319-10178-1.

- [6] T. M. van Engers and S. Nijssen, "Connecting People: Semantic-Conceptual Modeling for Laws and Regulations," in M. Janssen, H. J. Scholl, M. A. Wimmer, F. Bannister, editors, *Electronic Government, proceedings 13th IFIP WG 8.5 International Conference, EGOV 2014*, Springer, Dublin, Ireland, Sep. 2014, pp. 133-146, ISBN: 978-3-662-44425-2, e-ISBN: 978-3-662-44426-9.
- [7] T. M. van Engers, R. K. G. Winkels, and P. J. M. Kordelaar, "The know and the flow," in *Proceedings of the 5th International Symposium on the Management of Industrial and Corporate Knowledge*, Columbus, Ohio, United States, Oct. 2008
- [8] W. N. Hohfeld, "Fundamental Legal Conceptions as applied in judicial reasoning," Yale University Press, 1919.
- [9] G.W. Leibniz, "Dissertation on the Art of Combinations," 1666, in *Philosophical Papers and Letters, Part I*, 1989, pp. 73-84, ISBN: 978-90-277-0693-5, e-ISBN: 978-94-010-1426-7.
- [10] E. de Maat and R. Winkels, "Suggesting Model Fragments for Sentences in Dutch Laws," *Proceedings of Legal Ontologies and Artificial Intelligence Techniques*, May 2010 pp. 19-28 [Online]. Available from: <http://ssrn.com/abstract=2013146> 2015.01.10
- [11] E. de Maat, "Making sense of legal texts," PhD-thesis, Sep. 2012, ISBN 978 90 5335.
- [12] E. de Maat and T. M. van Engers. "Mission impossible?: Automated norm analysis of legal texts," in D. Bourcier, editor, *Jurix 2003: The Sixteenth Annual Conference, Legal Knowledge and Information Systems*, Amsterdam, IOS Press, Dec. 2003, pp. 143-144, ISBN: 978-1586033989.
- [13] M. Sergot, F. Sadri, R. Kowalski, F. Kriwaczek, P. Hammond and T. Cory, "The British Nationality Act as a Logic Program," in *Communications of the ACM*, Vol. 29, No. 5, May 1986, pp. 370-386, doi: 10.1145/5689.5920.
- [14] M. Sergot, "Representing legislation as logic programs," *Machine intelligence 11*, Oxford University Press, Inc., New York, NY, 1988, pp. 209-260, ISBN: 0-19-853718-2.
- [15] G. Sileno, A. Boer and T. M. van Engers, 2014, "Legal Knowledge Conveyed by Narratives: Towards a Representational Model," in M. A. Finlayson, J. C. Meister and E. G. Bruneau, editors, *2014 Workshop on Computational Models of Narrative (CMN)* Jul. 2014: pp. 182-191, doi: 10.4230/OASICS.CMN.2014.182.
- [16] G. Sileno, A. Boer and T. M. van Engers, "On the Interactional Meaning of Fundamental Legal Concepts," in *Legal knowledge and information systems: JURIX 2014: the twenty-seventh annual conference* Vol. 271. *Frontiers in artificial intelligence and applications*, pp. 39-48. IOS Press, doi: 10.3233/978-1-61499-468-8-39.
- [17] J. Wilkins, "An Essay Towards a Real Character, and a Philosophical Language," Gellibrand, 1668.

Study on Web Analytics Utilizing User Environment Segmentation in “Business to Business” site.

Akiyuki Sekiguchi

Graduate School of Systems and Information Engineering,
University of Tsukuba
Tokyo, Japan
sekichan2008@gmail.com

Kazuhiko Tsuda

Graduate School of Business Science,
University of Tsukuba
Tokyo, Japan
tsuda@gssm.otsuka.tsukuba.ac.jp

Abstract— In this study, we surveyed correlation of website access by user environment. Correlations can help us understand differences in user behavior that vary by time by place, or by user environment. We also found the user behavior tracking like visit numbers or page dwell time categorized by user segmentation is effective. For example, mobile device usage rate is higher in non-working hours than working hours and viewed pages are different. We confirmed user environment segments with a correlation approach can be used for web analytics for user navigation studies or even marketing use.

Keywords— web metrics; web analytics; B to B website

I. INTRODUCTION

Business-to-Business (B to B) manufacturer websites have changed in role and responsibility since use of the Internet has become widespread even among hardware engineers. The purpose of visiting a website has changed from searching for technical documents into searching for solutions or products without any face-to-face contact. In B to B manufacturer industry, traditionally sales related activity or even marketing related activity was tied to face-to-face salespersons' activity. In this period, the website's role was for searching and for providing technical documents or technical software resources. With the growth of e-commerce, users have become hesitant to meet salespeople and at the same time the manufacturer wants to track user behavior on the web and utilize analytic data for marketing and improving sales figures.

According to “The End of Solution Sales,” Harvard Business Review, July-August 2012, a recent Corporate Executive Board study of more than 1,400 B to B customers found that those customers completed, on average, nearly 60% of a typical purchasing decision—researching solutions, ranking options, setting requirements, benchmarking pricing, and so on—before even having a conversation with a supplier.“ This means many B to B customers select products/solutions and even buy them without intervention from salespeople. This means the website is becoming more important than before even from a business point of view even though it was used only for information delivery in the past. That is why we need study on dedicated B to B web site analytics.

We have been studying web metrics methodology and actual user behaviors just dedicated to B to B manufacturer

websites. In our past survey, there were two purposes for Business to Business (B to B) web analytics: (1) Improve and optimize the site for users by path analysis; and (2) Use in marketing activities. Compared to B to C web analytics, B to B web analytics have the following three characteristics:

(a) In many cases, the buyer is not the same person as the web user. So it is important to analyze all the users from the same company or organization as a single unit. That can be stated as B to B to C web analytics, not simply B to B.

(b) The goal of visitors to the website is often not only to make a purchase. Main conversions can be downloading a file, making an e-mail subscription or inquiring online.

(c) It is rare for a user to complete their goal within a single session. In most cases, users require multiple sessions spread out over a long period of time to complete their goal.

In past studies, we came up with proposals for B to B website analytics and also we studied the effectiveness of web analytics by user segmentation and especially the importance of page dwell time. In this study, we try to examine web analytics by user environment and see the effectiveness from this point of view. User environment means user access hours (time of day), connection type, and device type, etc. This study's data is based on web access data for a Japanese website from 2014 January 1st to September 30th from a global B to B manufacturing company.

Section II presents previous work. Section III shows user segmentation and focus metrics. Section IV is a study on visit times and dwell time by hours and connection type. Section V is web analysis findings by segmentation by content directory. Section VI shows analytics by web connected devices. Section VII is about major findings and we have conclusion in Section VIII.

II. PREVIOUS STUDIES

In our previous studies, we came up with a web analytics scheme for B to B websites and we defined B to B site conversion types and the importance of user registration on web. This is the first study of B to B type conversion related [16]. In another study, we checked effectiveness of page dwell time as well as traditional metrics like page views, unique users, visits and conversion rate. [15].

We found the importance of registered versus unregistered user segmentation and confirmed user behavior is different in each segment [17]. There are many studies on web metrics for e-commerce behavior [2] and a study on

personalization metrics on web [11], and studies on web metrics related to the B to B market [8]. There have also been general web analytic studies. [1][3]-[7][9][10][12]-[14].

III. USER SEGMENTATION AND FOCUS METRICS IN THIS STUDY

We categorized potential segmentation as shown in Table I. In the previous study, we saw content category segmentation and some others. In this study, we focus on connection type and the time period in which users are accessing the manufacturer’s website. We wanted to see the difference in behavior by user environment (Time and Place).

TABLE I. USER SEGMENT EXAMPLES

Segmentation category	Examples and considerations
1. By content category	Viewers of product information versus viewers of investment relations (IR)/company information User seeking to download software versus e-commerce users
2. By user environment (by time slot or by connection type)	By time hours Midnight users versus business hour users By connection type like through providers or through company network.
3. By user referrer	Users arriving through search engine, by e-mail click, or by bookmark/URL typing
4. By visit frequency	First time versus second and more frequent users
5. By user commitment level (registered or unregistered)	Registered users versus unregistered users
6. By company profile	Focus customer versus unfocused . Large customers versus small customers
7. By industry	User behavior by industry
8. By participation	Only web tracking for converted customers or unconverted customers
9. By device type	User navigation can differ by device.

IV. VISIT TIMES AND DWELL TIME BY HOURS AND CONNECTION TYPE

As a quick reference, for a B to B web site there is much difference in traffic between weekdays and weekends. However, the trends of traffic by time of day show almost the same peaks across all days. Figure 1 shows general trend in web visit number by hours.

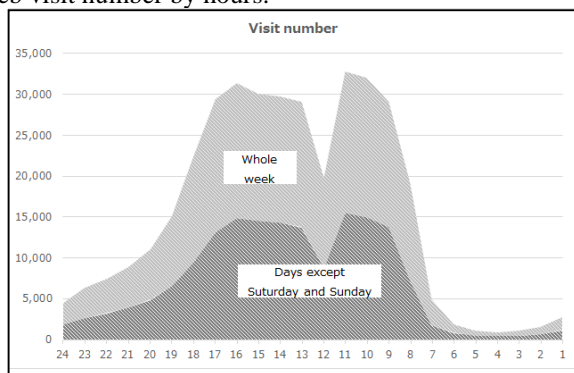


Figure 1. General trend in web visit number by hours

According to the statistics provided by the Japan Institute for Labor Policy and Training shown in [18], "hours of work per week, manufacturing" is 42.2 per week. Also, "9:00 AM to 7:00 PM" are typical working hours in Japan. We defined three time periods as "1. Home and commuting", "2. Work", and "3. Commuting and home". Firstly, we tracked user accesses by time period with consideration to company size. Normally small-sized companies or individual engineers tend to use normal internet providers and middle or large-scale companies use their own domains. We tracked them by time of day distinguishing the users who came through normal domains and users who came through internet providers (called "Providers"). Figure 2 shows a visit numbers trend.

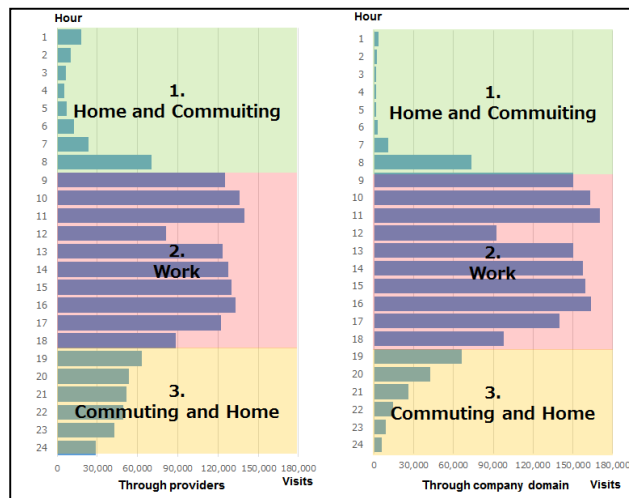


Figure 2. Visit numbers by time of day through providers and company domain

As a result the total coefficient of correlation of company and provider users is 0.59 and middle level of correlation (Similar trends) but if we omit the times from 1:00 AM to 9:00 AM, there is a strong correlation of 0.78. This means trends are similar throughout the day except for one time period. Only the time period between 1:00 AM and 9:00 AM shows some difference in numbers between “via providers” and “via company domain”. To illustrate this in more detail, Figure 3 shows visit numbers between 1:00 AM and 9:00 AM by hour. From late night to morning we assume some engineers work at home or work at small companies which use connections through providers.

This kind of data is useful for deciding which content should be shown or targeted to customers accessing the website at each time of day. Also, it can be assumed that provider users are made up of not just customers but also a general audience who are looking for IR information, company information or even some news through the sites. In fact during this time period numbers for these contents are relatively higher than during business hours. The proportion of IR/press release visits are 33% higher compared to normal working hours.

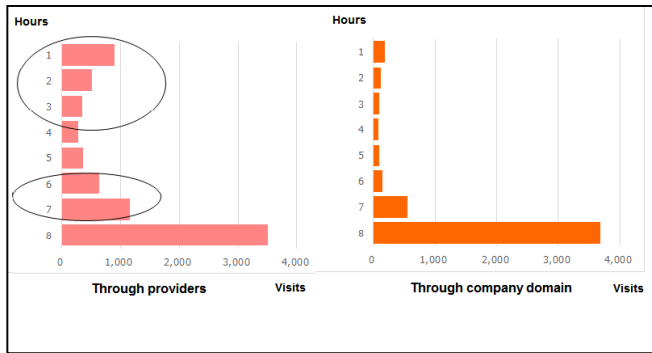


Figure 3. Visit numbers by time of day through providers and company domain from 1:00 AM to 9:00 AM

Next we looked at the page dwell time. Please refer to Figure 4. This is average of page dwell time by each hour for providers and company domain. The total coefficient of correlation is 0.68 for access via providers and via company domains. However, if we calculate the correlation for outside working hours, i.e. 7:00 PM to 1:00 AM, the correlation is 0.94. We can see some difference between the connection types for this hours. This shows the possibility of differences in usage between provider users and company users. The types of pages that are actually viewed are different. Generally, company users view purchasing information more and provider users view more press releases or IR information. We will investigate which information is accessed more in our next study.

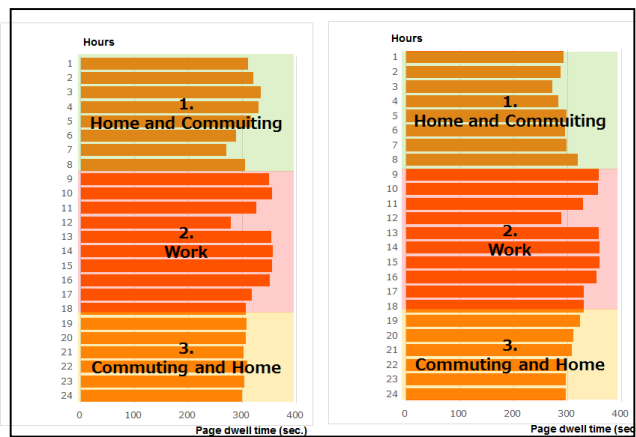


Figure 4. Page dwell time by time of day through providers and company domain

Please refer to Figure 5 for page dwell time by time of day by connection type. Page dwell time for small-sized customers who use providers peak at 10:00 PM and they probably work from home or on trains while commuting in Japan. This can be related to the fact that trains are the most common way of commuting in Japan. Also, for company

domain users this could indicate engineers working on development overnight.

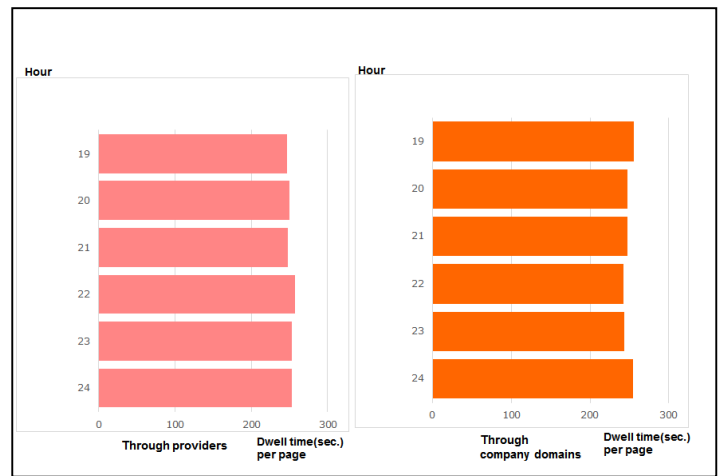


Figure 5. Page dwell time by time of day through providers and company domain from 7:00 PM to 1:00 AM

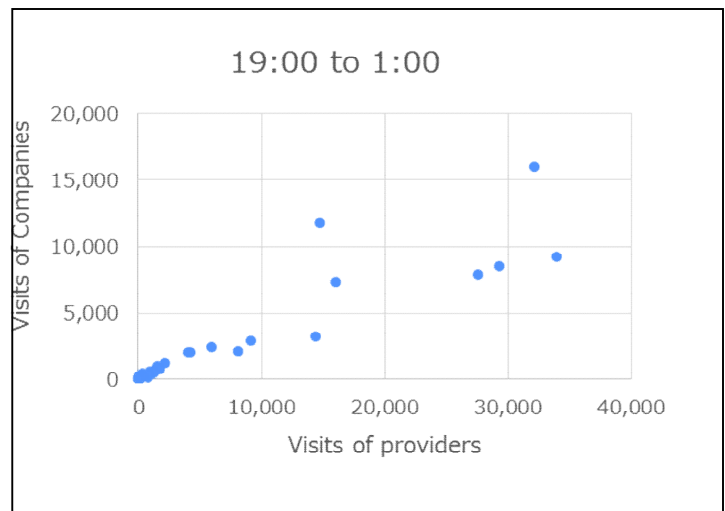


Figure 6. Correlation of visits from 19:00 to 1:00

Figure 6 shows correlation of visits from 19:00 to 1:00. Visit number is correlated between them and not su much difference.

V. ANALYSIS BY DIRECTORY

We surveyed the correlation between provider users and company users in terms of several segments. Firstly, we looked at correlation by content category (directory). There is a strong correlation for dwell time between providers and companies shown in TABLE II. TABLE II shows correlation coefficient in page dwell time for each hour and connection type. However, there is some different correlation just for some directories. Referring to TABLE III, the search function is one area of differentiation, especially in the 1:00

to 9:00 zone. Correlation here is lower than other periods. The number of searches performed by provider users is lower than those performed by company users. It is assumed that normally mobile access is through providers and these users are viewing websites during their train commute and do not search for any solutions or products but do view press releases or events during this time.

TABLE II. DWELL TIME CORRELATION TOTALLY

Directory Dwell Time		1:00 to 9:00		9:00-19:00		19:00to24:00	
		Providers	Companies	Providers	Companies	Providers	Companies
1:00 to 9:00	Providers	1.00					
	Company	0.84	1.00				
9:00-19:00	Providers	0.98	0.83	1.00			
	Company	0.88	0.92	0.88	1.00		
19:00to24:00	Providers	0.92	0.80	0.91	0.87	1.00	
	Company	0.91	0.84	0.90	0.86	0.91	1.00

Actually the company in this study tries to provide a different user interface to different customers depending on the time period. For 19:00 to 1:00, some navigation elements are changed with A/B testing, and the conversion rate for downloads is 125 times higher for time targeting. This type of analysis can be used for marketing purposes. For this purpose we will keep studying for further details.

VI. ANALYSIS BY DEVICE

We also looked at the relationship by device. We cannot see the actual device that a user owns but we can see information on OS (Operating System). There is much less correlation between providers and company trends in the time periods “1:00 AM to 9:00 AM” and “7:00 PM to 1:00 PM”. Most likely the mobile device usage rate is higher in non-working hours than working hours as shown in TABLE IV and TABLE V. Both table show correlation coefficient in page dwell time for each hour and connection type. Currently, unlike B to C sites the layout of most B to B sites is not mobile device compliant. However, B to B sites need to think about mobile device compliance especially for users who access through providers.

TABLE III. CORRELATION BY CONTENT DIRECTORY

Directory	1:00 to 9:00		9:00 to 18:00		19:00 to 1:00	
	Providers	Company	Providers	Company	Providers	Company
products	106,475	78,981	556,082	595,534	128,806	76,145
Search	21,377	30,217	167,138	193,837	35,397	22,751
support	26,514	19,350	120,251	127,775	32,221	15,962
press	20,321	11,382	89,733	77,086	34,035	9,186
comp	18,817	9,361	89,042	74,004	29,369	8,441
gur	16,049	16,175	84,245	115,158	14,808	11,761
edge_ol	15,972	6,116	64,111	52,709	27,620	7,830
applications	9,994	6,134	48,677	52,147	16,080	7,286
career	7,877	2,579	31,870	20,476	14,428	3,193
ir	6,894	3,747	25,322	21,894	9,203	2,854
disclaimers	4,852	3,736	17,557	17,190	4,247	2,021
company_info	4,164	1,663	18,134	14,629	8,118	2,064
event	4,131	3,638	18,706	22,214	4,125	1,997
partner	3,777	2,322	22,313	19,933	6,018	2,426
contact	2,238	1,821	12,022	11,829	2,163	1,257
public	1,602	1,758	9,243	12,234	1,627	960
buy	1,601	957	6,625	5,950	1,737	757
myrenesas	1,060	990	4,766	6,171	1,209	562
cmn	1,009	632	3,907	3,514	1,175	408
purposes	952	548	3,241	2,672	1,058	326
secret	906	908	6,164	7,962	973	587
videoclip	776	437	3,253	2,966	1,036	387
redirect	750	376	3,515	3,436	1,359	528
chat	640	439	3,580	3,678	613	338
Inquiry	557	495	3,023	3,145	436	371
search	459	564	2,396	3,644	406	433
user	436	133	1,981	829	833	133
edge	424	321	2,022	2,730	812	458
_print_this_page_	389	403	1,669	2,022	310	251
smart	311	171	1,399	1,485	453	173
devcon_jpn_2014	238	204	1,362	1,455	357	147
prod	147	135	959	1,192	196	167
ecology	131	69	461	613	198	103
media	128	182	681	1,413	153	271
facebook	103	41	548	277	187	46
sitemap	93	50	380	339	87	29
legal	88	75	383	461	104	46
tech	84	52	324	401	172	76
guidance	75	84	338	404	53	31
csr	63	20	231	157	115	25
privacy	44	95	212	270	57	18
campaign	34	24	151	137	55	14
lib	33	30	180	192	30	19
registration	32	25	116	139	18	15
rss	29	20	134	176	43	15
tool	29	23	116	220	24	22
C:	18	5	106	50	16	4
supp	15	8	40	86	30	8
r_video	13	9	80	102	30	6
manga	8	1	10	2	10	2

TABLE IV. CORRELATION BY OS TYPE

OS Type Dwell Time		1:00 to 9:00		9:00-19:00		19:00to1:00	
		Providers	Companies	Providers	Companies	Providers	Companies
1:00 to 9:00	Providers	1.00					
	Company	0.28	1.00				
9:00-19:00	Providers	0.97	0.27	1.00			
	Company	0.40	0.89	0.31	1.00		
19:00to1:00	Providers	0.91	0.33	0.91	0.39	1.00	
	Company	-0.03	-0.08	0.09	-0.15	-0.18	1.00

TABLE V. CORRELATION BY OS TYPE IN DETAILS

Item	1:00 to 9:00				9:00 to 19:00				19:00 to 1:00			
	Provider	%	Company	%	Provider	%	Company	%	ISP	%	Company	%
GNU/Linux	1,662	0.78%	632	0.43%	5,257	0.48%	4,152	0.35%	2,689	0.89%	1,121	0.77%
Microsoft Window:	162,070	75.83%	139,460	95.45%	947,855	85.98%	1,145,688	97.66%	211,999	69.86%	134,006	92.22%
Others	208	0.10%	19	0.01%	358	0.03%	91	0.01%	340	0.11%	46	0.03%
UNIX	38	0.02%	7	0.00%	104	0.01%	76	0.01%	47	0.02%	14	0.01%
Apple Macintosh	5,937	2.78%	1,710	1.17%	20,218	1.83%	9,893	0.84%	12,009	3.96%	3,028	2.08%
Unspecified	63	0.03%	16	0.01%	269	0.02%	96	0.01%	177	0.06%	34	0.02%
Google Android	19,312	9.04%	2,156	1.48%	55,271	5.01%	6,543	0.56%	33,457	11.03%	3,550	2.44%
Apple iOS	24,395	11.41%	2,091	1.43%	72,954	6.62%	6,504	0.55%	42,705	14.07%	3,493	2.40%
Microsoft Window:	24	0.01%	8	0.01%	52	0.00%	32	0.00%	18	0.01%	7	0.00%
Blackberry	7	0.00%	7	0.00%	20	0.00%	12	0.00%	14	0.00%	4	0.00%
Symbian	9	0.00%	3	0.00%	7	0.00%	0	0.00%	4	0.00%	1	0.00%
WebOS	0	0.00%	0	0.00%	1	0.00%	0	0.00%	0	0.00%	0	0.00%
Adobe	0	0.00%	0	0.00%	1	0.00%	1	0.00%	1	0.00%	0	0.00%

Other examples of no correlation are “viewed page numbers” and search usage time shown in TABLE VI and TABLE VII.

TABLE VI. VIEWED PAGE NUMBER CORRELATION

Viewed page numbers		1:00 to 9:00		9:00-19:00		19:00to1:00	
Dwell Time		Providers	Companies	Providers	Companies	Providers	Companies
1:00 to 9:00	Providers	1.00					
	Company	-0.23	1.00				
9:00-19:00	Providers	0.78	-0.33	1.00			
	Company	-0.30	0.48	-0.37	1.00		
19:00to1:00	Providers	0.68	-0.22	0.68	-0.30	1.00	
	Company	-0.17	0.34	-0.24	0.35	-0.14	1.00

TABLE VII. CORRELATION WITH SEARCH USAGE TIME

req (trend daily)		1:00 to 9:00		9:00-19:00		19:00to1:00	
Visit		Providers	Companies	Providers	Companies	Providers	Companies
1:00 to 9:00	Providers	1.00					
	Company	0.91	1.00				
9:00-19:00	Providers	0.70	0.87	1.00			
	Company	0.65	0.88	0.96	1.00		
19:00to 24:00	Providers	0.44	0.52	0.76	0.61	1.00	
	Company	0.54	0.79	0.92	0.94	0.70	1.00

There is much difference between time and connection type and there is possibility navigation can be further optimized according to time of day or connection type.

VII. SUMMARY OF FINDINGS

We found the user behavior tracking like visit numbers or page dwell time categorized by user segmentation is effective. Especially the accesses like time and place (or connection type) have different trend by each segments. For example the time period between 1:00 AM and 9:00 AM shows some level of difference in numbers between “via providers” and “via company domain”. Also, for page dwell

time 7:00 PM to 1:00 AM time period has differentiations between providers and companies. Also, depending on content type we found some difference. For example in the 1:00 to 9:00 zone user behavior is different and the number of searches performed by provider users is lower than those performed by company users. It is assumed that normally mobile access is through providers and these users are viewing websites during their train commute and do not search for any solutions or products but do view press releases or events during this time. We also looked at the relationship by device. We found that mobile device usage rate is higher in non-working hours than working hours and also viewed pages are different between them.

VIII. CONCLUSIONS

We have been trying to study the effectiveness of web analytics for a B to B manufacturer site with several studies. We defined some of the segment models and examined web access using some segments. In this study, we surveyed correlation of access by user environment. There are correlations between time of day or correlation between connection types such as connecting through a provider or through a company network. We used some key web metrics such as visits and page dwell time for our correlation survey. We noticed user environment segments with a correlation approach can be used for web analytics for user navigation studies or even marketing use. With the results of this study we will keep testing and doing analytics for targeted pages or targeted navigation by user environment as our next study.

REFERENCES

- [1] B. Adamson, M. Dixon, and N. Toman, “The End of Solution Sales Harvard Business Review” July-August 2012, pp102-106.
- [2] Y.Ichikawa, M Nakamura, Y.Kishimoto, and T.Kobayashi, “A Proposal of Extracting Innovative Users with Web Access Log of an E-Commerce site”, IPSJ SIG Notes 2012-GN-83(2); 2012, pp.1-8.

- [3] C. J. Aivalis, A. C. Boucouvalas, , Log File Analysis of E-commerce Systems in Rich Internet Web 2.0 Applications: This paper appears in: Informatics (PCI), 2011 15th Panhellenic Conference; 2011.p. 222-226.
- [4] J. Park, K. Jung, Y. Lee, G. Cho, J. Kim, and J. Koh, "The Continuous Service Usage Intention in the Web Analytics Services", System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on; 2009.pp. 1-7.
- [5] T. Ejiri, "Web Analytics and Web Marketing: Access Log Analytics Realized Web Marketing" Kaizen Cycle Management systems: a journal of Japan Industrial Management Association 18(1); 2008.pp. 8-43.
- [6] P. Sampath Dept. of Comput. Sci. & Eng., Bannari Amman Inst. of Technol., Sathyamangalam, India, "An efficient weighted rule mining for web logs using systolic tree", Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on; 2012.pp. 432-436.
- [7] S. Otsuka, M.Toyoda, and M. Kitsuregawa, "A Study for Analysis of Web Access Logs with Web Community", Information Processing Society of Japan (IPSJ), Database44; 2003.pp. 432-436.
- [8] A. Phippen, L. Sheppard, S. Furnell, "A practical evaluation of Web analytics", Internet Research, Vol. 14 Iss: 4; 2004.pp. 284-293.
- [9] K. Rebecca, P. Justin and P. "Graeme, Ethical considerations and guidelines in web analytics and digital marketing", a retail case study, Proceedings of the 6th Australian Institute of Computer Ethics conference 2012, Australian Institute of Computer Ethics, Melbourne, Vic; 2012.pp. 5-12.
- [10] X. Wang, D. Shen, H. Chen, L. Wedman, "Applying web analytics" in a K-12 resource inventory, Electronic Library, The, Vol. 29 Iss: 1; 2011. pp. 20-35.
- [11] W. Xiao-Gang, "Web mining based on user access patterns for web personalization", Computing, Communication, Control, and Management, 2009. CCCM 2009. ISECS International Colloquium; 2009.pp. 20-35.
- [12] G. K.New Mexico Inst. of Min. & Technol., Socorro, NM, USA Colbaugh, R., "Web Analytics for Security Informatics" , Intelligence and Security Informatics Conference (EISIC); 2011.pp. 20-35.
- [13] Pascual-Cid, V.Web Res. Group, Univ. Pompeu Fabra & Fundacio Barcelona Media, Barcelona, "An information visualisation system for the understanding of web data", Visual Analytics Science and Technology, VAST '08. IEEE Symposium; 2008.pp. 183-184.
- [14] N. O., L. Soliman, M., Saka, E., Badia, A, Germain, R., "A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites" , Knowledge and Data Engineering, IEEE Transactions; 2008, pp. 202-215.
- [15] A. Sekiguchi, K. Tsuda, "Consideration on page dwell time in B to B industry web analytics", ACIS; 2013. pp. 902-904.
- [16] A. Sekiguchi, T. Katsunuma, I. Hokao, Y. Yamada, K. Tsuda, "Web analytics for B to B marketing in semiconductor industry", International Journal of e-Education, e-Business, e-Management and e-Learning, Vol.2, No. 5; 2012.pp.413-418.
- [17] A. Sekiguchi and K. Tsuda, "Study on web analytics utilizing segmentation knowledge in business to business manufacturer site", KES 2014, pp.902-909.
- [18] "The Japan Institute for Labour Policy and Training", "Databook of International Labour Statistics", 2014, pp.1.

Tacit and Explicit Knowledge in Software Development Projects: Towards a Conceptual Framework for Analysis

Hanna Dreyer
The Business School
University of Gloucestershire
Cheltenham, UK
Dreyer.Hanna@gmail.com

Martin Wynn
School of Computing and Technology
University of Gloucestershire
Cheltenham, UK
MWynn@glos.ac.uk

Gerald Robin Bown
The Business School
University of Gloucestershire
Cheltenham, UK
GRBown@glos.ac.uk

Abstract – The management and delivery of software development projects remains a key business activity in many industries. Although the advent of packaged software products has reduced the incidence of in-house development, bespoke software is still important for some industrial sectors - notably in the finance, defence and security industries. Despite the recognized criticality of software project success for organizations, a considerable proportion of projects continue to either not meet their due dates, exceed budget, do not deliver to specification, miss quality targets, or do not meet customer requirements. Software project failure – be it bespoke products or the implementation of commercially available packages - remains an area of considerable interest in contemporary software project management literature, and the management and transfer of knowledge within both these types of project is a key dimension and driver of project outcomes. This paper examines how knowledge definition and management can be applied within a conceptual framework to improve software development project outcomes.

Keywords – software development; tacit knowledge; explicit knowledge; project management; conceptual model

I. INTRODUCTION

This research aims to uncover the connection between tacit and explicit knowledge in software development projects. McAfee [1] pointed out the dangers of differing interpretations of checklists in attempting to achieve success in software projects. He concluded that “the successful leadership of an IT implementation will continue to be a subtle craft”, in which the appropriate interpretation and application of knowledge are key. Tacit knowledge is difficult to articulate but is, according to Polanyi [2] [3], the root of all knowledge, which is then transformed into explicit, articulated knowledge. The process of tacit to explicit knowledge transformation is therefore a key component of software development projects.

This introductory section is followed by four further sections. Section 2 contains a discussion of the theoretical framework for this paper. Then, in Section 3, the research methodology is discussed. Section 4 outlines progress to date on developing a conceptual framework for analysis; and finally, Section 5 looks at how this research can be further progressed.

II. THEORETICAL FRAMEWORK

Knowledge is conceived of as a collection of information or usable data that is put into context, in order to comprehend how something works. Unlike most goods, knowledge does not have a physical form and cannot therefore be touched. The question remains as to what extent it can be spoken. The relationship of the whole, the entire focal point, to the parts, the detail, is a partnership where one part has larger intensity than the other [3]. Developing Polanyi’s work, Nonaka and Takeuchi [4] suggest that tacit knowledge can be converted into explicit knowledge, and therefore expressed in words or numbers. Human knowledge transfer is one of the greatest challenges in contemporary society, due to knowledge often being inaccessible. Its initial creation, transfer, utilization, and storage are often problematic, both conceptually and operationally.

Nonaka and his co-authors’ studies on knowledge, although influential, still leave a number of unanswered issues, such as those concerned with group tacit knowledge and explicit knowledge [5]. There have been many critics of tacit knowledge since Polanyi’s Personal Knowledge was published in 1962 [6]. Tacit knowledge is personal knowledge that an actor knows he has, but which he cannot describe in terms other than its own performance. An opinion is that “there can be no doubt that Personal Knowledge comes at us with its rhetoric all out of focus. The lack of clarity can be said to be due to its heterology; “it is a mixed bag” [7]. Critically challenging linguistic based knowledge with the idea that “we can know more than we can tell” [2], tacit knowledge still needs more exploring in order to be verified. These elements provide the rationale for this research and define the space it seeks to address.

Knowledge creation is a continuous, self-transcending process by which one can transcend the boundary of the old self into a new self by acquiring a new context, a new view of the world and new knowledge [8]. The raw product of knowledge is information, which differs greatly from knowledge due to it not being involved with a specific context. “Without context, it is just information, not knowledge” [8]. Having a lot of information will not serve a company unless it has been evaluated, put into context and agreed that it is a justified true belief.

From individual knowing, the development of shared understanding is important in any joint enterprise. Knowledge is a key asset in companies; it can serve as a

basis to build a greater knowledge base and can therefore provide a competitive advantage [8]. In differentiating between tangible assets, which can only be used by one party, and intangible knowledge assets, which can be shared and used by several parties, the utilization, transfer and storage of knowledge becomes a core issue.

Social network analysis plays a significant role when analyzing the process of knowledge creation within a group. Some authors [9] have focused on ways network analysis can be used to identify key players in diffusion networks. "Diffusion of innovations looks at the process by which a new technology or idea gets adopted by a given community" [9]. Furthermore, Prell [10] assesses ego networks, an individual's network, and the influence it has on their alter-egos. These tools aim to help explore the interactions of project members and their positioning within the group in order to evaluate knowledge inputs and outputs.

The environment of this research is a software development project, which influences tacit and explicit knowledge through the time restricted character of projects; the clearly defined beginning and ending [6]. As previously stated, human knowledge transfer is one of the greatest challenges in today's society; not having sufficient time to articulate tacit and explicit knowledge affects project success. In addition, work is allocated and distributed throughout the different actors of the project, which results in the scattering of knowledge throughout project members.

III. METHODOLOGY

The research sets out to identify the impact of tacit and explicit knowledge transferred during software development projects. An inductive, exploratory, qualitative methodology is being applied in order to validate the tacit knowledge spectrum in software development projects. The philosophical foundation of this study is based on the ontology of subjectivism, while the epistemological position is interpretivism. The researcher is centrally involved in the phenomena being studied, and is a key player in the process of data collection and analysis to answer the research questions. In terms of the methodological approach, a case study method is adopted (see Figure 1); this is deemed appropriate given the embedded nature of the study. If a case study strategy incorporates multiple cases, then the resulting data can provide greater confidence in the research findings [11]. Data will be generated through unstructured interviews, and will therefore be accessed in a narrative form [8]. Personal reflection on the data collected during the meetings, as well as participant observation, will also be part of the methodology. The study sets out to evaluate three software development projects as case studies in different companies.

Seeing the company as a contextualized culture, and the participant observation as performative auto ethnography [12], will aid the conduct of this inductive research. Knowledge is assessed by textual forms, empathetic epistemology of critical and co-present reflection with others, transformation to the dominating system, and social as well as linguistic effects with others. This will result in a contextualized conceptual model, proposed for the analysis of tacit and explicit knowledge in software projects.

As tacit knowledge does not have a physical form, the information passed on throughout the project needs to be interpreted in a qualitative form. Generating data through unstructured interviews, participant observation and personal experience will help identify the tacit knowledge within the data. Pre-interviews with the key players in the project, as well as short interviews to track any change of view, and a final conversation, will be part of the interview data generation process.

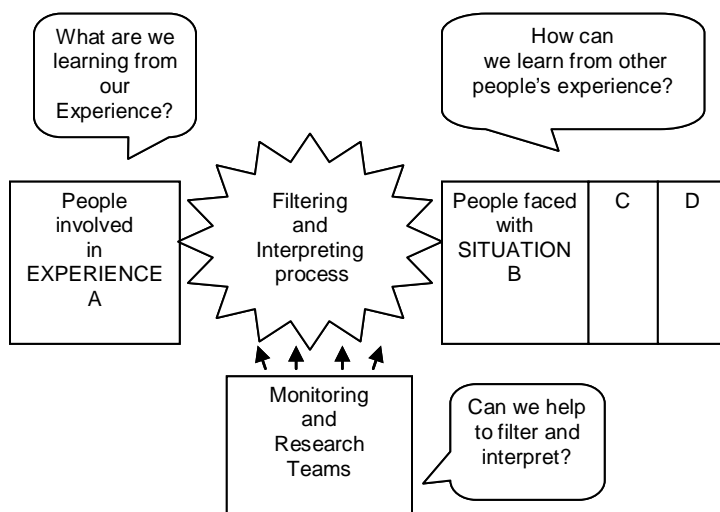


Figure 1. Case Study Research [13]

The aim is to find information in the data that will be collected in over 30 hours of meetings from the first case study. Following this, the results are to be refined through two more software development projects in other companies, which will each consist of 5 hours of meetings. Follow up meetings with several project members will be used to validate the data, and thereby develop the tacit knowledge spectrum in software development projects. This will be done through the analysis and interpretation of the data collected during these meetings. Software development meetings are recorded and transcribed according to the importance of the information; an assessment of the exchanged tacit knowledge in relation to the tacit knowledge spectrum will be produced. The aim of the follow up meetings is to confirm the themes revealed in the data gathered during the meetings. Validation, elaboration, as well as evaluation of the previously found themes should result. Finding key players, and therefore the key knowledge sources, through participant observation will support the evaluated data from the meetings. This should, in the future, aid researchers as well as managers to understand and further develop tacit knowledge in the work place, and in particular in software development projects.

In summary, the main elements of the research method and design are:

1. Qualitative exploratory research
2. Inductive research

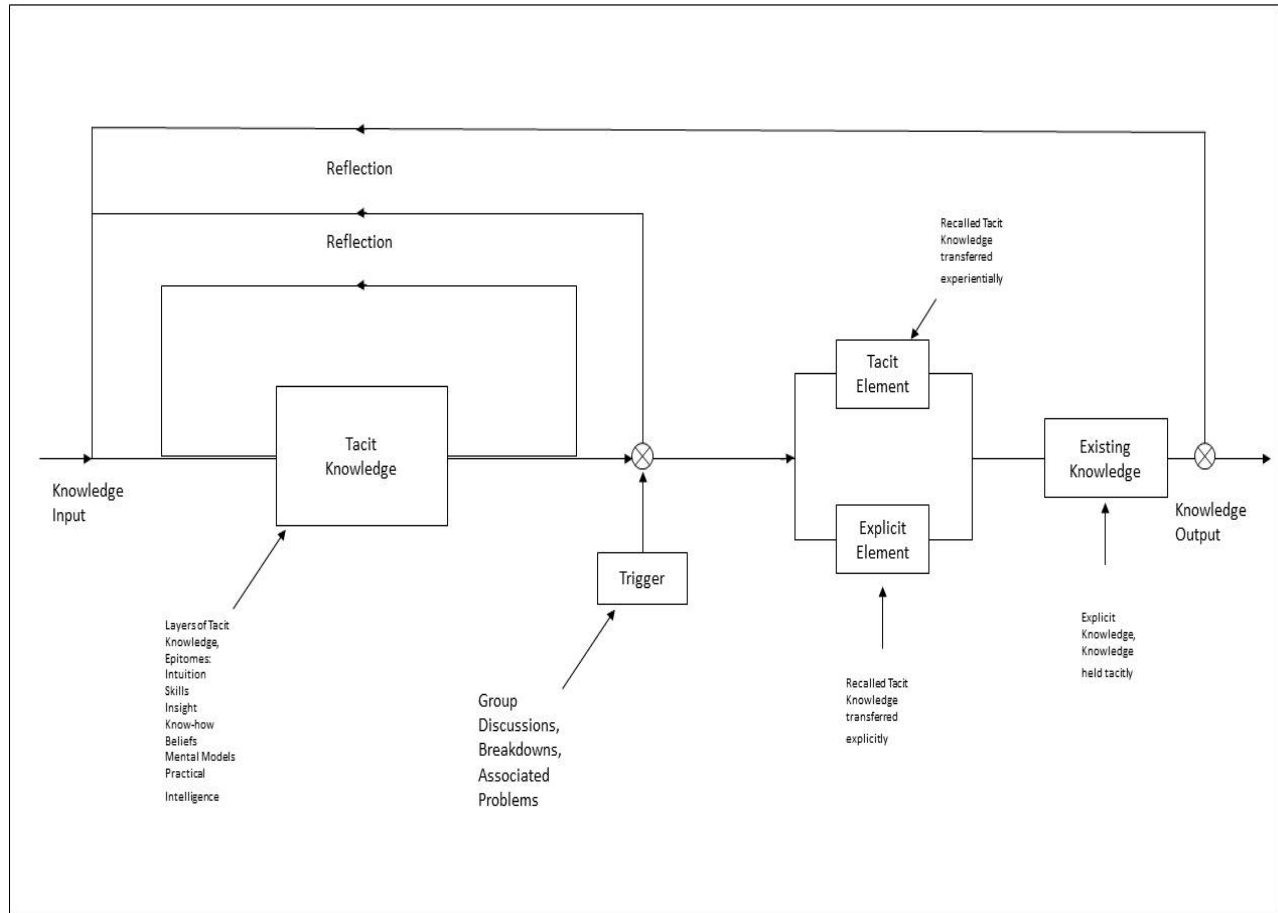


Figure 2. The Tacit Knowledge Spectrum [15]

3. Participant observation
4. Personal reflection
5. Unstructured interviews

Bias will be eliminated through personal reflection and constant validation of the data gathered. Once the emergent themes are found, the interviews will be constructed and later held with the project members. Interviews with the project participants will be held, in order to validate the interpretations and data garnered from the meetings. Through replication logic, external validity will be ensured; internal validity is not needed in exploratory research.

IV. CONCEPTUAL FRAMEWORK DEVELOPMENT

Most of the research in the field of software development projects is quantitative and does not take into account the qualitative inductive aspects of the topic. The main focus of the study is to qualitatively understand the transfer, as well as the impact, of tacit to explicit knowledge and its effect on the development of new software. The aim is to create a model that supports future software development projects in their tacit to explicit knowledge transfers.

A clearly defined beginning and ending, as well as the work to be done, are the main trademarks of projects. The Office of Government Commerce defines a project as “a temporary organization that is needed to produce a unique and pre-defined outcome or result, at a pre-specified time, using predetermined resources” [14]. Therefore, projects are time restricted to a larger extent than day-to-day business. This directly affects software projects since data generation has to be done within the scope of the project and cannot be repeated, due to their unique and time restricted nature. Project failure or delay is often due to poor communication, which is closely related to the exchange of knowledge. Another challenge of projects is the storage and communication of knowledge resources during and after completion. With the advent of email and the internet, knowledge is now mainly transferred through electronic devices in information technology projects, which has greatly changed the information flow. This influences the impact of tacit and explicit knowledge, which, due to it being unarticulated, often remains hidden. In addition, project environments are dismantled, and knowledge resources are often lost once a project is completed. These issues demonstrate the specific character of projects and their long

term benefits, in particular when they concern software development.

The recent work of Clarke [15] puts forward a model of a tacit knowledge spectrum developed from a study of three companies (see Figure 2). Using Clarke's model to test tacit and explicit knowledge provides a provisional conceptual framework to help the research analysis process. Approaching theory as "a way of seeing and thinking about the world rather than an abstract representation of it" [16] will set the data into perspective. Using Clarke's model, and the underlying theory of tacit knowledge, creates the opportunity to investigate theory in the work place. This is based on the three functions of directing attention, organizing experience and enabling useful responses [16] [17].

Clarke's model will help to sensitise and explore the use of tacit and explicit knowledge in software development projects. The research aims to contribute knowledge in the area of knowledge management by increasing awareness about the tacit and explicit knowledge transferred, and its impact on such projects. Given the time and budget constraints normally imposed on software development projects, the goal is to aid project managers and team members in their future planning, as well as improve the transfer of knowledge and information during a project. Through an analysis of existing literature, allied to empirical data and observations in large project environments, this research will look to develop Clarke's conceptual framework and answer the following research questions:

- What is the current understanding of knowledge exchange in software development projects?
- How can tacit and explicit knowledge be recognised and evaluated in software development projects?
- To what extent does non-communicated tacit and explicit knowledge amongst team members influence the project and its acceptance?
- Can tacit and explicit knowledge be better harnessed through the development of a conceptual model for use in software development projects?

This approach assumes that it is feasible and sensible to cumulate findings and generalize results to create new knowledge. The concepts of tacit and explicit knowledge will be analysed in primary research case studies. The key assumption that there is a "trigger", that acts as a catalyst for the recall and transfer of different knowledge elements, will be examined in software development projects. The basic conceptual framework will build on Clarke's Tacit Knowledge Spectrum and will be developed further in the light of further literature analysis and first-hand project research.

V. CONCLUDING REMARKS

Peter Drucker used to tell his students that when intelligent, moral, and rational people make decisions that appear inexplicable, it's because they see a reality different to the one seen by others [18]. This observation by one of the leading lights of modern management science underscores the importance of knowledge perception and knowledge

transfer. With regard to software projects, McAfee [1] noted that "the coordination, managerial oversight and marshaling of resources needed to implement these systems make a change effort like no other". Yet, although software project successes and failures have been analysed within a range of analytical frameworks, few studies have focused on knowledge transfer. This phenomenon requires further research into the interaction and communication of knowledge within and between project teams and their contexts, and this is the purpose of this research, in the specific context of software development. If knowledge can be more successfully harnessed to improve the software development process, it has the potential to significantly enhance eventual project outcomes.

REFERENCES

- [1] A. McAfee, "When too much IT knowledge is a dangerous thing," MIT Sloane Management Review, Winter, 2003 pp. 83-89.
- [2] M. Polanyi. The tacit dimension. The University of Chicago Press, 1966.
- [3] M. Polanyi, Personal Knowledge: Towards a Post-Critical Philosophy. Routledge, 1998.
- [4] I. Nonaka and H. Takeuchi, The knowledge-creating company: How Japanese companies create the dynamics of innovation. Oxford: Oxford University Press, 1995.
- [5] Z. Erden, G. von Krogh, and I. Nonaka, The quality of group tacit knowledge, 2008
- [6] M. Polanyi., Personal Knowledge, Psychology Press, 1962.
- [7] T. Langford and W. Poteat, "Upon first sitting down to read Personal Knowledge: an introduction". In Intellect and hope: Essays in the thought of Michael Polanyi, 1968, pp. 3-18.
- [8] I. Nonaka and D. Teece, Managing Industrial Knowledge, 2001, London: SAGE.
- [9] T. W. Valente and R. Davies, "Accelerating the diffusion of innovations using opinion leaders", The Annals of the American Academy of Political and Social Science, vol. 566, 1999, pp. 55-67.
- [10] C. Prell, Social Network Analysis: History, Theory and Methodology, 2012, London: SAGE
- [11] R.K. Yin, Case Study Research (4th Edn), Sage, 2009
- [12] D. S. Madison, Critical Ethnography: Method, Ethics and Performance, Second Edition 2012, London: SAGE..
- [13] M. Wynn, J.L. Taylor, and J. Overall, "Case study monitoring: an approach to urban management training", Planning and Administration, vol. 9, no. 1, Spring 1982, pp 16-24.
- [14] Office of Government Commerce (OGC), Managing Successful Projects With PRINCE2, 2009, London: The Stationery Office/Tso.
- [15] T. Clarke, The development of a tacit knowledge spectrum based on the interrelationships between tacit and explicit knowledge. 2010, Cardiff: UWIC.
- [16] A. Deetz, Doing Critical Management Research, 2000, London: Sage.
- [17] M. Alvesson and K. Skoeldberg, Reflexive Methodology, 2004, United Kingdom: SAGE.
- [18] B. Baker, , "The fall of the firefly: An assessment of a failed project strategy," Project Management Journal, vol. 33, no. 3, 2002, pp. 53-57.

Knowledge Processes in German Hospitals

First Findings from the Network for Health Services Research Metropolitan Region Bremen-Oldenburg

Lars Rölker-Denker, Insa Seeger, Andreas Hein

Department of Health Services Research
University of Oldenburg
Oldenburg, Germany
lars.roelker-denker@uni-oldenburg.de
insa.seeger@uni-oldenburg.de
andreas.hein@uni-oldenburg.de

Abstract—This paper summarises the work on organisational and network learning process among hospitals in Germany. It is a descriptive analysis of learning processes identified by interviews with board of directors and observations in clinical wards. A shift from organisational learning to inter-organisational/network learning could be confirmed. This paper mentions some of them and gives an outlook on further work.

Keywords-health services research; organisational learning; inter-organisational learning; hospitals; Germany.

I. INTRODUCTION

The increase of knowledge and information is a general phenomenon and thus also applies to healthcare. Emerging cooperation between health care organisations (HCO) and in addition, Mergers & Acquisitions by highly integrated health care groups extends the organisational knowledge base even more and results in inter-organisational cooperation. In addition, medical schools and medical university hospitals represent key actors in medical knowledge development [1].

Organisational learning routines are key factors for learning HCO like hospitals in general [2] and clinical wards [3] but also for larger network structures [4]. This paper will present some newly identified organisational and inter-organisational learning routines from field studies in two running projects in German hospitals.

In section 2, both projects, the methodology and data assessment are briefly introduced. Section 3 contains the results so far, separated into organisational and inter-organisational/network level. Section 4 discusses the weaknesses of the approach and the comparison with the state of the art, especially Lipshitz/Popper [3]. The paper closes with conclusion and outlook in Section 5.

II. METHODOLOGY AND DATA ASSESSMENT

The project “Network on Health Services Research in the Metropolitan Region Bremen-Oldenburg” started on January 1st 2014 with duration of 2 years [5]. A framework for semi-structured interviews has been developed on several health

related topics like cooperation, gaps in healthcare provision and knowledge processes. The knowledge process related questions are:

- *Organisational level*: Are learning routines established between the clinical wards?
- *Inter-organisational/network level*: Do these learning routines exist between the hospital and other health care organisations like family doctors, rehabilitation centres and other?
- What is the role of information and communication technology for these learning processes?
- What is the role of learning processes for patient care?

Due to the design of the project, which has as a strategic task the initiation of a (research) network, it was decided to conduct interviews instead of performing a survey. For networking reasons it was important to get in personal contact with the board of directors. In addition, the researcher aimed at raising the answering rate by personal contact. Between April and October 2014 data from 22 hospitals have been gathered, by only having 3 hospitals with negative feedback not willing to take part in the study. Interviewed persons have been from the board of directors, mostly chief executives and medical directors, sometimes with the nursing director.

In addition, there are observations and interviews in context of a doctoral study. Three HCOs (two hospitals and one rehabilitation centre) are analysed by 1-2 day-observation of clinical daily routine, supplemented by interviews with personnel from the observed wards. Up to now, seven wards have been observed and four interviews have been conducted. Additional information on methodology and results on the doctoral study can be found at [6] and [7].

III. RESULTS

In this section the identified learning processes are listed and, if applicable further literature is indicated. They are distinguished between organisational and inter-organisational level. Learning processes on the

organisational level only refer to processes inside of the health care organisations. Inter-organisational learning processes occur when at least one professional health actor from outside the health care organisation is involved.

A. Organisational level

The following learning processes have been identified on the organisational level:

- *Morning/lunch conferences*: these interdisciplinary conferences take place one or more times a week and bring together physicians from the general disciplines of internal medicine or surgery. These conferences are across clinical wards, e.g., in the internal medicine conference physicians from all sub-disciplines like cardiology, geriatrics, nephrology and others meet and discuss interdisciplinary or organisational issues. Sometimes these meetings are enriched by small presentations on a specific medical problem [7].
- *eLearning applications*: eLearning applications from different sources with focus on physicians (CME – continuous medical education) or nursing are also part of the organisational level domain inside a hospital (e.g., for Germany see [8]).
- *Conventions/closed-door meetings*: these kinds of meetings also occur in hospitals in different manifestations. One complete ward (e.g., the cardiological ward) or one professional group (e.g., all surgeons' assistants) come together for one or two-day meeting abroad and discuss strategic issues for the future and review the past period. This meeting can also be used in integrated healthcare groups (inter-organisational/network level), e.g., all geriatric physicians of all hospitals from one healthcare group meet once a year.
- *Interdisciplinary case conferences/clinical conferences*: these conferences are organised e.g., once a week and difficult cases are discussed in an interdisciplinary matter [7]. They can be distinguished between internal medicine and surgery, but also conferences between these general disciplines are possible. The typical process starts with a physician having a difficult case and signs up the patient to the clinical conference. During the conference the patient's case is discussed involving all necessary data and information (x-ray, laboratory results, etc.) and in the end, a recommendation is given to the enquiring physicians.
- *Radiological conference*: Clinical radiological conferences are a modification of the interdisciplinary case conference. During this conference the radiological ward demonstrates the recent imaging (x-ray, x-ray computed tomography, magnetic resonance imaging) and discusses the results with the clinical wards, e.g., with the neurosurgical ward on neoplasms located in the brain and how to proceed.

- *Interdisciplinary therapeutically pathways/clinical pathways*: These pathways are derived from medical guidelines (developed by medical societies) and adapted to the unique clinical context of a hospital. The adaption and introduction process is performed by interdisciplinary team of physicians and nurses; afterwards an interdisciplinary monitoring team controls the execution and initiates slight adjustments [9].
- *M&M conference (mortality and morbidity conference)*: these meetings aim on retrospective analysis of patient's data. The focus can be on medical errors, as well as on the teaching value [3] [10].
- *Idea management*: Established idea management processes are also implemented in healthcare organisations like hospitals. As in other branches they are an important factor for organisational improvements.
- *Critical Incident Reporting Systems (CIRS)*: CIRS are used for reporting critical (nonlethal) incidents in an anonymous manner. The anonymous way of reporting is intended to raise the rate of reporting incidents and protecting the reporting person [11].
- *Surgical operation reflection meetings*: these meetings aim on interaction and reflection between physicians (surgeons, anaesthetist) and other personnel involved in surgical operations like surgical nurses or surgeon's assistants [3].

B. Inter-organisational/network level

The following learning processes have been identified on the inter-organisational/network level:

- *Conventions/closed-door meetings*: similar to conventions/closed-door meetings on organisational level (see Section III.A.).
- *Tumour conferences/tumour boards*: Tumour boards refer to oncological diseases and cancer and are often carried out as interdisciplinary meetings. They connect clinical oncologists with residential physicians (family doctors and oncologists) in order to improve the compliance of the patient and guarantee ongoing care after hospital treatment [12]. In rural areas with a low physician and/or population density, these meetings can be carried as remote meetings supported by tele- or videoconferences and/or electronic medical records [13].
- *Rehabilitation conferences*: these conferences are similar to tumour boards but focus on rehabilitation. In Germany, clinical care/hospitals and rehabilitation are separated into two different sectors. For example, clinical geriatricians and rehabilitation geriatricians, in some cases complement by residential physicians, meet one a regular basis and discuss patients moving from the clinical sector to the rehabilitation sector.
- *Diseases-related regional networks*: many healthcare organisations are active in disease-related networks, often with a regional focus. These

networks differ from hospital networks, mixed hospital-family physician networks and networks of family physicians.

- *Advanced education with network organisations*: usually the hospitals offer their advanced education programme to their employees exclusively. In this case, selected parts of the programme are also offered to resident physicians, e.g., regular radiological trainings for clinical and resident radiologists or nursing trainings for multiple hospitals.
- *Tele-diagnosis*: tele-diagnosis is a tele-medical application and refers to remote diagnosis. This is often used in rural areas or areas with low density of medical specialist or hospitals [14].
- *Common Quality management with network partners*: quality management (QM) is an established process in hospitals for many years. These QM systems can also be expanded to network partners like ambulatory health centres connected to a hospital.
- *Common ward conferences*: in integrated healthcare groups regular ward conferences can be expanded to all group hospitals, possibly scaled by regional or other parameters, e.g., a group-wide conference of all neurological wards.
- *Common leadership of clinical wards among several hospitals*: in integrated healthcare groups several hospitals may have one head physician for e.g., all geriatric wards.
- *Network CIRS*: hospital-wide CIRS (see III.A) can be expanded to all hospitals of one healthcare group or to other connected organisations like ambulatory health centres.

IV. DISCUSSION

A. Comparison with state of the art

Lipshitz and Popper [3] have identified eleven organisational learning processes in hospital environments (they call them organisational learning mechanisms) in their early work. Table I gives an overview on the processes identified by Lipshitz/Popper and the processes verified during the interviews, including the frequency of occurrence. The occurrence is wrapped up by the participating hospitals and not by the individual interviewees.

In addition, the following organisational learning routines have been identified during the interviews:

- *Interdisciplinary case conferences/clinical conferences*: one hospital, also validated by interviews in two clinical wards.
- *Interdisciplinary therapeutically pathways/clinical pathways*: one hospital also validated by interviews in one clinical ward.
- *Radiological conference*: two hospitals, also validated by observations in three surgical wards.
- *Idea management*: one hospital.

The organisational learning processes on an inter-organisational level had not been focussed by Lipshitz/Popper in the past. In particular, the inter-organisational processes identified by interviews are listed in Table II.

All other inter-organisational learning processes have been only mentioned once, this applies to:

- Conventions/closed-door meetings
- Rehabilitation conferences
- Advanced education with network organisations
- Tele diagnosis
- Common Quality management with network partners
- Common ward conferences
- Common leadership of clinical wards among several hospitals
- Network CIRS

Li et. al. [19] performed a systematic review on communities of practice and also identified learning processes on an organisational and inter-organisational level. In particular these are fostered interaction between students and expert physicians, informal learning clubs (both on an organisational level), health care agency collaboration, and virtual communities over organisational borders (both on an inter-organisational level). All these processes have been identified by the interviews.

B. Weaknesses/Gaps

The interviews focussed on the board of directors. This implies some weaknesses:

- The medical directors are from discipline, either one surgical oriented or one internal medicine discipline. This mind-set has an influence on the discussion results, e.g., a surgeon will more focus on surgical reflection meetings. Interdisciplinary process should be focussed by both disciplines.
- There might be more learning processes in the distinct wards not being visible to the medical directors in their role as a board member, e.g., there are some specific geriatric processes like the multi-professional geriatric team session [7]. For a complete overview, the interviews have to be conducted with at least all chief physicians of the relevant wards. Since the project focusses on a first overview this approach is acceptable.
- There might be also learning processes in the distinct wards which are not visible to the board members in their role as managers in their wards. This can be coped with observations in the wards which are an approach in the associated doctoral studies [7]. In addition, there are some learning processes being naturally covered by the daily work in the wards, e.g., daily physician rounds, and, due to this, not being mentioned by the board of directors.

TABLE I. OCCURRENCE OF ORGANISATIONAL LEARNING PROCESSES

Organisational Learning Process	Frequency of occurrence/remarks
Physicians' rounds	Naturally covered by daily work, validated in all wards by observations
Staff meetings	Naturally covered by daily work, validated in all wards by observations
Reviews of medical records	Naturally covered by daily work, validated in all wards by observations
Nursing staff meetings	3 hospitals
Journal club	2 hospitals, also validated in all wards by interviews and observations
Morbidity-mortality conferences	1 hospital
Reflections in and after surgery	1 hospital
Periodic reviews	Not mentioned
Research reports	Not mentioned
Clinical pathological conferences	Not mentioned
Video demonstrations	Not mentioned

TABLE II. OCCURRENCE OF INTER-ORGANISATIONAL LEARNING PROCESSES

Inter-organisational Learning Process	Frequency of occurrence/remarks
Tumor conferences/tumor boards	5 hospitals, also validated by interviews in two clinical wards
Diseases-related regional networks	3 hospitals (adiposity, trauma, tumour surgery)
Common head physicians rounds of clinical wards among several hospitals	2 hospitals

- Both approaches (interviews with chief physicians and observations) have a bigger demand on time and personnel resources than semi-structured interviews with the board of directors. For practical reasons a good balance between resources available and study objects to analyse needs to be defined. In the associated doctoral studies the observations and interviews with chief physicians are focussed on two hospitals and one rehabilitation centre and, since the studies have a focus on geriatrics, on the geriatric wards and the wards with the most important patient movements.

V. CONCLUSION/OUTLOOK

A. Conclusion

Organisational learning in healthcare has evolved from an intra-organisational phenomenon to an inter-organisational/network phenomenon in Germany. This is similar to developments in other countries. For instance, since the mid of the 1990s cancer networks have been emerging and fostering the inter-organisational exchange between involved physicians [15].

Another reason for emerging inter-organisational learning processes is enhanced cooperation between hospitals for economic reasons; on one hand more cooperation between hospitals on a regional level, on the other hand more integrated healthcare groups (e.g., Helios or Asklepios in Germany). In these settings, intra-organisational learning processes have been adapted to the new network environment, e.g., CIRS originally used in distinct hospital settings and now used in healthcare groups.

B. Outlook

The interviews in the project Network on Health Services Research in the Metropolitan Region Bremen-Oldenburg will be continued, same applies to the work in context of the doctoral studies.

Next step is the modelling of selected learning processes with a combined approach of 3LGM² (Three-Layer Graph-based Meta Model; developed for clinical IT architectures and hospital functions modelling [16]) and KMDL[®] (Knowledge Modelling and Description Language [17]; based on the knowledge conversion of Nonaka & Takeuchi [18]) which has already started [7]. Based on these future findings the velocity of knowledge dissemination will be measured and, if necessary, process remodelling proposals will be suggested. A best-practice transfer is also planned among the participating hospitals.

ACKNOWLEDGMENT

This publication is partly funded by the Metropolregion Bremen-Oldenburg (application number: 23-03-13).

REFERENCES

- [1] L. Rölker-Denker and A. Hein, „Learning hospitals from health services research perspective. Study design and method inventory“, DMW, 2012, vol.137, A281, doi:10.1055/s-0032-1323444 [German: Lernende Krankenhäuser aus versorgungsforschender Perspektive. Studiendesign und Methodeninventar]
- [2] H. Pfaff, „The learning hospital“, Z Gesundheitswiss, 1997, vol.5, pp. 323-342 [German: Das lernende Krankenhaus]
- [3] R. Lipshitz and M. Popper, “Organizational Learning in a Hospital,” J Appl Behav Sci, 2000, vol.36, pp.345-361, doi:10.1177/0021886300363005
- [4] L. Rölker-Denker, “Hospitals as Learning Organizations”. Proc. IADIS Int Conf e-Health, 2010, pp.295-298

- [5] L. Rölker-Denker, I. Seeger, and A. Hein, „Project presentation „Network on health services research metropolitan region Bremen – Oldenburg““, *Z Palliativ Med*, 2014, vol.15, p.15, doi:10.1055/s-0034-1374113 [German: Projektvorstellung „Netzwerk Versorgungsforschung Metropolregion Bremen – Oldenburg“]
- [6] L. Rölker-Denker and A. Hein, „Organisational learning routines in acute geriatric care and rehabilitation. Results of a qualitative study“, *Z Palliativ Med*, 2014, vol.15, p.117, doi:10.1055/s-0034-1374490 [German: Organisationale Lernroutinen in der geriatrischen Akutbehandlung und Rehabilitation. Ergebnisse einer qualitativen Studie]
- [7] L. Rölker-Denker and A. Hein, „Knowledge Process Models in Health Care Organisations. Ideal-typical Examples from the Field“, *Proc Healthinf* 2015, pp.312-317
- [8] J.T. Dilling and S. Bohnet-Joschko, “Integrated E-Learning in hospitals,” in *Wissensmanagement im Krankenhaus. Effizienz- und Qualitätssteigerungen durch versorgungsorientierte Organisation von Wissen und Prozessen*, S. Bohnet-Joschko, Eds. Springer, pp.63-77, 2008 [German: Integriertes E-Learning in Krankenhäusern]
- [9] L. de Bleser, R. Depreitere, K. de Waele, K. Vanhaecht, V. Vlayen, and W. Sermeus, „Defining Pathways,“ *J Nurse Manage*, 2006, vol.14, pp.553-563, PMID:1700496
- [10] J.D. Orlander and B.G. Fincke, “Morbidity and Mortality Conference. A Survey of Academic Internal Medicine Departments,” *J Gen Intern Med*, 2003, vol.18, pp.656-658, doi:10.1046/j.1525-1497.2003.20824.x
- [11] R. P. Mahajan, “Critical incident reporting and learning,” *Brit J Anaesth*, 2010, vol.105, pp.69-75, doi:10.1093/bja/aeq133
- [12] N.L. Keating, M.B. Landrum, E.B. Lamont, S.R. Bozeman, L.N. Shulham, and B.J. McNeil, “Tumor Boards and the Quality of Cancer Care”, *J Natl Can Inst*, 2013, vol.105, pp.113-121, doi:10.1093/jnci/djs502
- [13] C.L. Marshall, et. al, “Implementation of a Regional Virtual Tumor Board: A Prospective Study Evaluating Feasibility and Provider Acceptance”, *Telem e-Health*, 2014, vol.20, pp.705-711, doi:10.1089/tmj.2013.0320
- [14] ATMA: American Telemedicine Association. What is Telemedicine? [online]. Available from <http://www.americantelemed.org/about-telemedicine/what-is-telemedicine/> 2014.10.26
- [15] R. Addicott, G. McGivern, and E. Ferlie, “Networks, Organizational Learning and Knowledge Management: NHS Cancer Networks”, *Public Money and Management*, 2006, vol. 26, pp.87-94
- [16] A. Winter, B. Brigl, G. Funkat, A. Häber, O. Heller, and T. Wendt, „3LGM²-modeling to support management of health information systems“, *Int J Med Inform*, 2007, vol.76, pp.145-150, PMID:16962819
- [17] N. Gronau, „Modeling and Analyzing knowledge intensive business processes with KMDL: Comprehensive insights into theory and practice“, GITO, 2012
- [18] I. Nonaka and H. Takeuchi, „The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation“, 1995, Oxford University Press
- [19] L. C. Li, et. al. „Use of communities of practice in business and health care sectors: A systematic review“, *Implementation Science*, 2009, vol.27, doi: 10.1186/1748-5908-4-27

S-Grouper

A Semantic Based System to Semi-Automatic Encode Hospital Activities

Roberta Cuel and Andrea Francesconi
 Department of Economics and management
 University of Trento
 Trento, Italy
 {roberta.cuel; andrea.francesconi}@unitn.it

Filippo Nardelli
 Expert System
 Trento, Italy
 fnarderlli@expertsystem.it

Giampaolo Armellin
 Centro Ricerche GPI
 Trento, Italy
 Giampaolo.armellin@gpi.it

Abstract— This paper presents a pilot implementation of a semantic based system which enables physicians to semi-automatic encode hospital cases. This system, called S-Grouper, is the result of a project carried out by the University of Trento and two Italian companies: GPI S.p.A and ExpertSystem. S-Grouper is aimed to improve the codification process of the so called Diagnosis-related group by (i) implementing semantic content analysis of healthcare records, (ii) providing useful hints and coherent categories that physicians may select, (iii) comparing the codification provided by professionals with the one automatically extracted from the patients' health care records, and (iv) improving management of knowledge.

Keywords- *Diagnosis-related group; coding processes; semantic analysis; information systems; healthcare quality.*

I. INTRODUCTION

In Italy, as in many other countries, hospital acute activities are classified, measured and financed according a standard classification system called Diagnosis-related group (DRG) [1]. DRG is identified manually or semi-automatically by physicians when the treatment on a patient is finished and the discharge letter is written. This activity is often supported by a 3M tool called Grouper that identify the DRG code analyzing the list of ICD-IX-CM codes which represent both diagnosis and clinical procedures adopted in the treatments.

We have analyzed various practices in different hospitals in Italy, and in all cases physicians have to analyze the patients' clinical records, manually codify diagnosis and clinical procedures, identifying the related ICD-IX-CM codes. Then the Grouper tool identifies the corresponding DRG code.

As in a socio-technical system [3], the process of codification (ICD-9-CM and DRG) is interrelated with the individuals' roles in the organization, knowledge they share, and their motivations. It is not hard to imagine, physicians do care more on medical procedures than on administrative

duties, they consider the processes of code identification as irrelevant and oppressive, they may not pay attention on the correctness of the codes, and provide results prone of mistakes [2].

Analyzing the working environment, procedures, inner interests of physicians, and the technology they use, we identified some recommendation to improve the DRG codification process [4, 5, 6].

II. OUR METHODOLOGY

Our approach encourages the interdependencies between social and technical sub-systems and the relations among users, tasks they carry on, technology they use and the social structure they belong to [3]. Our focus does not reflect a disregard for the technical aspects of software engineering, but is meant to underline features of the process that are oftentimes neglected by software developers, but are essential for the success of any socio-technical based application. Ideally the process of design and development starts with a field analysis aimed at identifying the motivations of individuals and the groups' practices which they belong to. Direct observations, interviews and questionnaires are very effective techniques that can be used to unveil and better define behaviors and motivations. In the second phase, the raw knowledge is then analyzed and requirements identified. The third phase is the creation of the software prototype which should be the simplest possible solution that can effectively support the users. In the fourth phase, the resulting prototype is tested, initially in a controlled environment with selected testers, then with real users, tasks (daily activities that actors usually carry on), and situations (the field and the social structure which actors belong to). The software changes in response to these findings, and the process is repeated until the desired outcome is achieved.

III. A SEMANTIC BASED TOOL

After a first analysis of clinicians practices we defined the requirements for a new tool aimed at improve the ICD-

IX-CM and DRG codification processes. The resulting tool called S/Grouper had the following high level architecture (Figure 1).

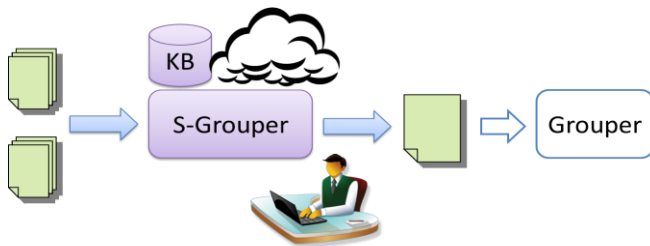


Figure 1. The system architecture of S-Grouper

S-Grouper is an expert system that:

- integrates semantic linguistic analysis in the Grouper software
- analyzes medical records and any other relevant document the physicians may need,
- and suggests some ICD-IX-CM categories.

These are used by physicians to:

- encode the medical treatments, reducing the time they spend dealing with bureaucratic procedures,
- suggest alternative codifications enabling clinicians to improve their abilities,
- check the coherence between the treatments described in the healthcare records and the codification chosen by doctors.

- analyze the document and extract entities from unstructured sources of information (see the underlined text at the right) such as the patients' health care records, the resignation letters, the nursing and surgical registers, etc.
- identify various ICD-IX-CM codes and propose them to the physicians (box at the left)
- select, organize and change the ICD-IX-DM codes (box at the left)
- recognize the most appropriate DRG code and close the patient discharge letter.

From the back office side, the system provides useful knowledge that enable administrative officers to:

- compare the resulting codes provided by the semantic engine with the ones manually listed by doctors in the hospital discharge form
- identify some lacks in physicians abilities
- measure the performances of physicians and identify some opportunistic behaviors
- report on potential errors or omissions, and provide some alerts
- measure the benefits generated by a more accurate reporting and ICD-IX-CM identification
- reduce time in dealing with bureaucratic procedures.

Finally, the public administration or the local government may take advantage of this tool. With S-Grouper they can automatically compare the DRG codes identified by physicians with the ones automatically identified through the semantic analysis of the patient's health care records, identifying opportunistic behaviors in DRG reporting..

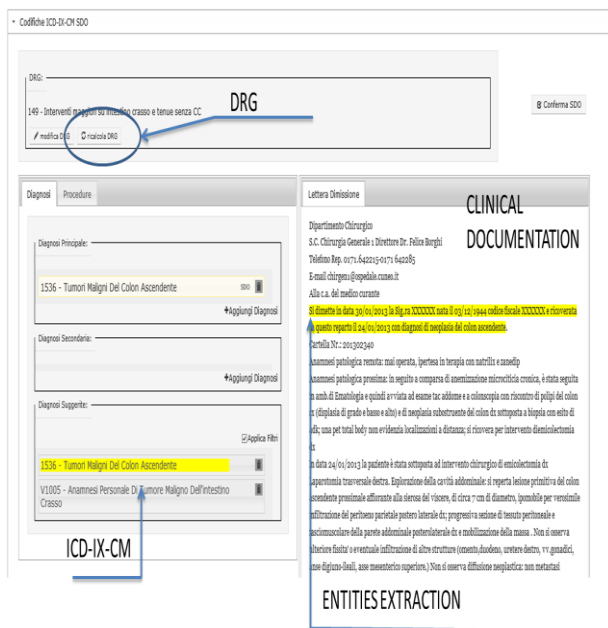


Figure 2. The interface of S-Grouper

As shown in the Figure 2, S-Grouper allows physicians to:

- connect in a "transparent" mode to the clinical documentations already digitalized (box at the right)

REFERENCES

- [1] Barr, Charles E., et al. "Conceptual Modeling for the Unified Medical Language System." Proceedings/the... Annual Symposium on Computer Application [sic] in Medical Care. American Medical Informatics Association, 1988.
- [2] Bowker, Geoffrey C. & Susan Leigh Star. *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: The MIT Press, 1999.
- [3] Bostrom, Robert, Saurabh Gupta, and Dominic Thomas. A Meta-Theory for Understanding Information Systems Within Sociotechnical Systems. *J. Manage. Inf. Syst.* 26, 1 (July 2009), 17-48.
- [4] Hoelzer, Simon, Ralf K. Schweiger, and Joachim Dudeck. "Transparent ICD and DRG coding using information technology: Linking and associating information sources with the extensible markup language." *Journal of the American Medical Informatics Association* 10.5 (2003), 463-469.
- [5] Pelzer, M., et al. "Electronic documentation of injuries of the hand with a semantic network: effective and efficient methods for the documentation of clinical and administrative processes." *Der Unfallchirurg* 110.3 (2007): 213-218.
- [6] Straub, Hans Rudolf, and Michael Lehmann. "A semantic clinical data repository-how the work on DRGs can serve clinical medicine too." *Swiss Medical Informatics* 27.71 (2011), 34-36.

ChoreMAP: Extracting And Displaying Visual Database Summaries Tool

I. Cherni

LTSIRS, Institut des Sciences
de Gestion of Tunis, Tunisie,
LIRIS, Institut National des
Sciences Appliquées de Lyon,
France

email: c.ibtissem@hotmail.fr

S. Faiz

ISAMM, de la Manouba,
Tunisie
LTSIRS, ENIT de Tunis

email: sami.faiz@insat.rnu.tn

R. Laurini

LIRIS, INSA of LYON,
Villeurbanne, Lyon, France

email: robert.laurini@insa-
lyon.fr

M. Warghi

Faculté des Sciences
Juridiques, Economiques et
de Gestion de l'Université de
Jendouba, Tunisie

email: mariem2012.warghi@
gmail.com

Abstract- Traditional cartography is an essential tool to describe the facts and the relations concerning a territory. Expert users are usually satisfied with the expressive power of traditional mapping, when it deals with simple cases. But in some complex cases including a large number of data, the expert users need a map which stresses the most important aspects rather than have several maps with a high level of details. It is in this context that our search has been launched in order to automatically discover spatial patterns and view based on a spatial database and chorems. This paper presents the project focusing on the extraction subsystem of salient patterns that will be encoded with an extensible markup (XML)-based language called chorem markup language (ChorML) and then be viewed as visual summaries by visualization subsystem.

Keywords- Chorem, geographic databases, summaries, geographic knowledge, data mining.

I. INTRODUCTION

The geographic databases contain the necessary information to the understanding of our environment; thus they help us make decisions related to our environment. Of course this information needs a clear and easy representation to be understood. Therefore, to be able to make decisions we need maps that give a synthetic vision of a whole which integrates visual summaries that easily describe and explain the important information.

Thanks to chorems, we are able to represent the knowledge we have about a certain place in a very simple and clear way. It is also thanks to their abilities to symbolize and encapsulate a methodology and its corresponding interpretation. We can also show climatic, geographic, economic, geologic, and agronomic situations, based on their spatial and statistical temporal context, by combining many chorems to form a chromatic card.

In this paper, we describe the research carried out within an international project, launched among several research institutions. The project is meant to define cartographic solutions able to better represent geographic information extracted from (spatial) database contents, which refer to both static objects and evolutionary phenomena.

It is by studying chorems that we were motivated to contribute to the development of a system to represent situations, like those mentioned previously, starting by the results of data mining on a geographic database. The most common situations are those represented in the study of the structure and the dynamics of population, urban concentrations or the interaction between natural and social systems. In these models, there is a problem to define a method to generate chorems that algorithms used to accede a database then extract patterns, which are visualized on a chorematic card.

The objective of this paper is to present the challenges related to the automatic extraction of the most significant elements, and then the creation of the chorems from the geographic databases.

The paper is organized as follows: in the next Section, we propose the model that we study in this paper, and some definitions of chorems are given. Section 3 presents clearly the problem. In Section 4, the architecture of our system is presented. Section 5 presents the extraction sub-system. In Section 6, the computational simulation is made to illustrate the efficiency of the algorithm. Finally, we conclude the paper with a case study in Section 7.

II. RELATED WORKS

An immediate synthesis of data of interest, disregarding details, is a real support for human activity to model and analyze the reality of interest. Such a synthesis may then represent the starting point for further processing tasks aimed at deriving spatial analysis data and represent them in a clear map so that expert users could transfer the principal idea of the problem in a simple way or could satisfy a decision maker. In the remainder of this Section, we describe two cartographic representations: cartograms and chorems. According to the definitions in [7], a cartogram is "a small diagram, on the face of a map, showing quantitative information" or "an abstracted and simplified map the base of which is not true to scale".

While cartograms show values of a single variable at a time, chorems allow designers to assemble into a single map more than one thematic layer, thus, representing the relative importance of a set of objects and phenomena related to

each other. According to the definition of the French geographer Roger Brunet, who invented chorems [1], a chorem is a schematized territory representation, which eliminates any detail not necessary to the map comprehension. In this definition, the term schematized means that the most important characteristic is a sort of synthetic global vision emphasizing salient aspects. Moreover, Brunet defines the chorems as “elementary structures of the space represented by graphic model used so that all the spatial configuration rise from the combination, eventually very complex, of simple mechanisms“ [2].

Now, a second definition of chorem can be proposed: “a chorem can be seen as a visual way to represent geographic knowledge, and so it can be a tool to summarize geographic databases” [5].

Nowadays, there are many systems to create cartograms automatically like QGIS cartogram creator, Scape Toad, Arc Toolbox, etc.,. But, there is just one system that is created to generate chorems [11]. ChEViS (Chorem Extraction and Visualization System) is an international project lunched in 2009. It is composed of two subsystems: chorem Extraction subsystem and Chorem Visualization subsystem. This system proposes a new method to construct chorems automatically, especially chorems for each user.

III. PROBLEM STATEMENT

Expert users are usually satisfied with the expressive power of traditional mapping when it deals with simple cases. But in some complex cases including a large number of data, the expert users need a map which stresses the most important aspects rather than having several maps with a high level of details.

So, our objective is to define geo-visualization solutions which can adequately represent the information extracted from geographic data. Visual models based on chorems can interpret and represent spaces, their geographic distributions and their dynamics. The same space can be represented in different ways, but all the corresponding maps will tell the same thing. We cannot change the message, its position, its hierarchy, its network, and all those items expressed in the chorematic map. So in this paper, we answer many questions, such as:

- How can we build chorems from geographical databases?
- Which algorithm is used to elaborate clustering to generate chorems?
- How can we extract knowledge to represent the spatial interaction between clusters?

IV. SOLUTION

This paper is issued from several experimentations concerning chorems, “visual schematic representation of territories”. The project is an international one. It was developed jointly by several research laboratories. It aims to define mapping solutions capable of synthesizing the

content of geographic databases and represent them in a plain, legible and intuitive way.

Figure 1 shows the architecture ChoreMAP (Chorem Mapping) [14], which is composed of three sub-systems: two extraction subsystems and a display subsystem. The previous figure shows the architecture of the proposed system, which consists of three main levels at both extraction and visualization subsystems.

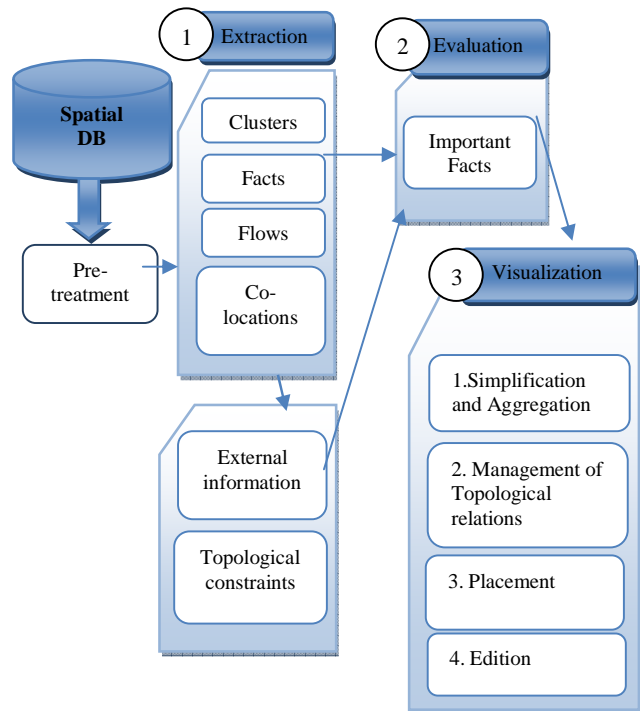


Figure 1. General architecture of the ChoreMAP prototype.

The extraction patterns subsystem is invented to obtain and manipulate information from available data. The extraction of important chorem subsystem manipulates the data generated by the first subsystem. It was strengthened by the integration of space technology of data mining, filtering based on SQL; as to the procedure to reduce the number of patterns, it requires the intervention of an expert taking into account what he wants to show on his card. Thus, the visualization subsystem manages this information by assigning a visual representation in terms of chorems and chorematic cards. It is important to note that the chorem structure is managed by the three subsystems. However, they use and/ or modify the various subparts of chorem structure. In particular, the two extraction subsystems manage chorems conceptual properties, such as name, type, size and coordinated by expressing them by alphanumeric attributes. The sub-display system develops and modifies the geometric shapes of chorems from conceptual properties generated by the previous phase in order to assign an appropriate visual representation. Communication between

levels is based on the multi-level ChorML language. In particular, a suitable level of ChorML is used to correspond to different specifications and to adapt to different useful formats in the running process.

V. SPATIAL PATTERN DISCOVERY

As indicated, the architecture is decomposed into three parts: chorem extraction, chorem important extraction and chorem visualization (Figure 1). The application of data mining and spatial data mining algorithms will result in patterns that will be later associated with chorems.

The Chorem Extraction System first transforms the actual data base of available geographic data in order to facilitate the extraction of significant information by spatial data mining. Then, Chorem Visualization System manages this information by assigning them a visual representation in terms of chorems and chorematic maps [5]. Once the patterns are extracted, these results are encoded in the ChorML language by a special subsystem for generating ChorML documents [3], and then visualized.

As previously stated, since there are four types of patterns resulting of the data mining, which seem to be the most interesting in the discovery of chorems, our extraction subsystem is composed of four parts:

- Subsystem of extraction facts.
- Subsystem of extraction flows.
- Subsystem of extraction of clusters.
- Subsystem of extraction co-localization patterns.

In order to encode this knowledge, a special language named ChorML was designed [3]. This language is composed of three important instances: First, ChorML 0 shows the database where data mining techniques are going to be applied. At this point spatial data is encoded in GML. Second, ChorML 1 is defined to hold the list of extracted patterns resulted from data mining processes. It contains some additional information such as the elements in the vicinity of the territory, a description of the boundary run on a list of topological and structural relationships between patterns, and finally, at ChorML 2 Geometric shapes have passed through a generalization process and coordinates are translated into a layout format – SVG [4].

A. Facts

A fact is considered the result of one or more queries against the database [9]. A set of rules is defined in order to obtain basic information from the database [4]. To achieve this sub-system, two methods are required: one to analyze data through a user request and another to encode and store the results in a file ChorML.

B. Clusters

Clustering is the method used to group data into classes. Consequently, an object in a cluster has certain similarities with other objects in the same cluster. For example, we could group parcels in a city or by their land use type or group regions by their weather similarities [9]. Clusters are

strong candidates to generate chorems [11]. There are many different algorithms for clustering (for example the distribution of clusters). But which one do we use to realize these spots (or subsystem)?.

After a detailed study of methods of data mining, we set the most appropriate algorithm k-means clustering to group the cities that are geographically close and share common characteristics of a set of groups fixed in advance. The proposed method is composed of three modules. It is represented in Figure 2.

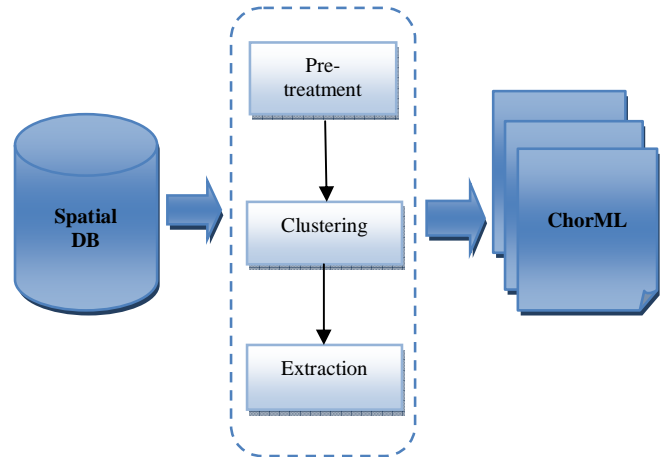


Figure 2. Architecture of sub-system of the clustering extraction

Pre-treatment module: The geographic data are different to those traditional ones. They are stored in the base in thematic layers. In our work we only consider digital data like longitude and latitude to implement the grouping. This module is equipped to create a file with extension “Arff” from data stored in the database.

Clustering module: In this module, we are interested in gathering cities into a set of clusters. In our work, by selecting areas that have common characteristics, we find that the spatial clustering and more specifically the Kmeans algorithm is the most appropriate method. This method allows to extract knowledge that match the pattern structure in the cluster ChorML language with Euclidean distance, Kmeans allows us to select areas with a maximum set of connected points that share the same characteristics. We use weka to affect extraction using Kmeans. Our algorithm is developed:

Step1: Enter the number of clusters (k).

Step2: Choose arbitrary k cities. Affect these cities in the k clusters (cities are the centers of the clusters).

Step 3: Assign each city V cluster C_i M_i center as $dist(V, E)$ is minimal.

Step 4: Recalculate M_i of each cluster (the geometric center).

Step 5: go to step 3 if we just made an assignment.

Extraction module: This module generates a document of clusters from the knowledge extracted by the selection and representation module.

C. Flows

One study showed that three types of flows are the most important [4]: flow path, divergent source flow and well oriented flow. The stream type represents a flow path where the origin and destination are well defined and it may possibly have a geometric shape (for example a large arrow). While the divergent flow type source has a definite origin, the destination is a little uncertain. The destination is a list of different geographical directions. Finally, the convergence has a definite destination, but the origin is a list of convergent geographic directions.

Flows are used to represent the spatial dynamics within a territory. "We consider as flows every movement, material or immaterial, of goods, of people, of information, between different locations"[8]. Flows are generally represented by arrows in the current mapping. We are particularly interested in the flow of goods.

Several authors, like Yann et al. [13] and Tobler [12], find that the flow of goods is mainly related to three factors: (i) the emissivity of the zone i (ii), the attractiveness of zone j and (iii) the inverse function of the distance between the two zones i and j. The distance between the zones has a space factor and an economic separation. Based on these studies, to retrieve related knowledge flows between clusters, we propose a method to study the available quantity of goods. Through the census figures of the city, the consumption and production of products for each city, we can get a good approximation of the movement of goods. This method is done by comparing the production and consumption of agricultural products.

First, to extract the flow of goods, it is necessary to study the quantities available in each city for each product. This method is to subtract the quantities produced: a product with the quantities consumed of the same product. The quantity calculation equation is:

$$Q_t = P_t - P_i * C_i \quad (1)$$

Or

P_t = population in year "t"

C_i = consumption in the year "t" of a product (p) for a person living in the city (V)

P_t = production of a product for the city (V) during the year "t"

Q_t = the quantity demanded (missing) for the product (p) in year t for the city (V).

After the preprocessing module where we store the geometrical shapes in the thematic layers and the descriptive data in a cube data base, we apply the proposed method of goods flow extraction. This method consists of:

- Calculating the quantities of goods available for all cities in the database (using Talend Open Studio).
- Determining the quantities available for each product group from the cluster extraction subsystem.
- Identifying clusters that have seized a quantity above the threshold. These clusters are considered as emitter groups of the product.
- For each receiver group, we compare the amount available for the quantities available for group issuers and the distance between the groups. The selected cluster is the nearest cluster and has an available quantity for the product p.
- Encoding flows in ChorML language.

D. The co-location patterns

Co-location patterns are sets of characteristics of places that are presumed to be close with a certain probability to each other. Co-localization rules seem interesting in creating chorems because they define the organization of objects within the territory with a quantitative accuracy. The results of the extraction modules and those made of clusters are used to determine the relationship between major cities to the user and the resulting groups of the k-means algorithm. For example, if there are commonalities between two geometric shapes, we have a relation of the type 'Touch'. The boundaries of these shapes touch, but their inside does not.

VI. EXPERIMENTS

The inland transport of merchandise is carried by road, rail or inland waterway. According to international definitions, transportation means a flow of goods moved over a given distance and is measured in ton-kilometers.

According to [15] and given the large amount of digital information available in the world, statisticians have the difficult task of ensuring that the trade analysts and others have speed access to accurate business data.

Accurate and up to date production is costly and requires resources that are unfortunately still lacking in many developing countries. That is why the frequency and level of detail of national statistics vary considerably from one country to another. It was often difficult to establish timely and comparable statistics on trade in some of developing countries because these countries do not regularly communicate.

Demographic and economic dynamics of major Tunisian cities is an undeniable reality in the recent reconfiguration of the Tunisian territory. As a basic source of the Tunisian economy, intra-regional trade in food products is considered as the main flow of goods.

Taking into account their importance, and because of the difficulty of their expert analysis, we find it interesting to address this limitation and propose an easier method. In fact, it would be interesting to represent flows on a chorematic map. This provides an easy and synthetic vision as massive

data rate in the territory and donated goods will be replaced by forms and symbols easy to understand.

In what follows, we turn to the testing phase where we use as input system a ChorML file from ChoreMAP project [14] that contains the flow of food and agricultural products between regions

A. Generation of chorematic map

In this Section, we present the five main features of our system which are:

- Extraction of the possible groups according to the number entered by the user.
- Extraction of cities issuing a selected product.
- Extraction of the flows of goods between different groups.
- Creating the ChorML document.
- Test results obtained by display on a chorematic map.

We present in the following the results of our experimental study. When we apply the extraction flow method on goods imported from the National Institute of Statistics (INS) in Tunisia and the Regional Offices of Agricultural Development (CRDA) actual data, the result of the experiment is a chorematic map presented in Figure 3, which shows its performance in the presentation of the flow of grain between the main regions in Tunisia.

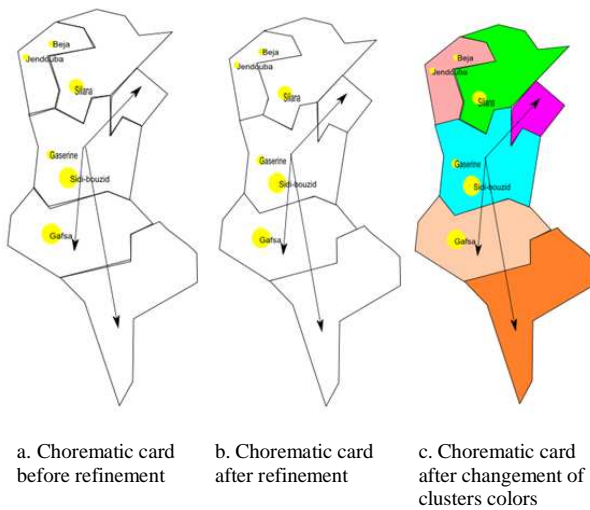


Figure 3. Tunisian Chorems of flows of cereals.

This Chorematic map was generated by using the results obtained from the extraction phase and knowledge encoded ChorML. It describes the quantity transported from one region to another taking into account the proposed threshold.

The menu offers an overview of the most salient freight elements process. It can be used as a decision support. This

map shows more precisely chorematic directions grain transportation in Tunisia. It consists of three chorems:

- The territory is divided into six major regions: Northwest, Northeast, Central east, Central west, South east, and Southwest.
- The most productive cereals cities in Tunisia.
- Arrows representing the flow of inter -regional cereals.

With this chorematic card, we can see the existence of the important movement of grain between Tunisian areas. Indeed, the Center west supplies other regions (South east, South west and Center east).

Regions producing cereals are characterized by the relative importance of the rural population and the high proportion of agricultural employment in total employment.

VII. CONCLUSION

Traditional cartography is an essential tool to describe the facts and the relations concerning a territory. Geographic concepts are associated with geographic symbols and graphic symbols help the readers understand immediately the visualized data.

The representation of chorems provides us with the best interpretation of problems. It is in this way that we can obtain all what we need: from the young pupils who want to learn geography, up to the researchers who investigate new forms of communication.

If we want to process other types of information without being confused, we propose the concept of chorem layers to present various phenomena and we also offer the superposition.

In this paper, we present our project ChoreMAP that allows us to define solution maps and display the information extracted from a geographical data base.

Future work can be summarized in the following activities:

- Conduct comparative experimental studies to verify the effectiveness of the proposed method. In particular, such studies aim to measure the ability of the expert users to exploit chorematic maps produced by the system.
- Check chorems can be used as a navigation tool or as a query language and access to the content of geographic databases.

REFERENCES

[1] R. Brunet, "La carte-modèle et les chorèmes", Mappemonde 86/4, 1986, pp. 4-6.
 [2] R. Brunet, "Les fondements scientifiques de la chorématique", In "La démarche chorématique", Centre d'Études Géographiques de l'Univ. Jules Verne, 1993.
 [3] I. Cherni, K. Lopez, R. Laurini and S. Faiz, "ChorML: résumés visuels de bases de données géographiques", International conference on Spatial Analysis and GEomatics, Paris, France, 2010, pp. 691-692.

- [4] A. Coimbra, "ChorML: XML Extension for Modeling Visual Summaries of Geographic Databases Based on Chorems", Master Dissertation, INSA-Lyon, France, 2008.
- [5] V. Del Fatto et al., "Potentialities of Chorems as Visual Summaries of Spatial Databases Contents", Springer Verlag LNCS, 4781, 2007, pp. 537-548.
- [6] V. Del Fatto, "Visual Summaries of Geographic Databases by Chorems", thesis INSA de Lyon, 2009.
- [7] D. Dorling, "Cartograms for human geography", Visualization in Geographical Information Systems, 1994, pp. 85-102.
- [8] M. Egenhofer, "A Formal Definition of Binary Topological Relationships". In: Foundations of Data Organization and Algorithms, 1989, pp 457-472.
- [9] Z. Guo, S. Zhou, Z. Xu, and A. Zhou, "G2ST: a novel method to transform GML to SVG". In: 11th ACM international symposium on Advances in Geographic Information Systems, Association for Computing Machinery, 2003, pp 161-168.
- [10] J. Han, M. Kamber and A.Tung, "Spatial clustering methods in data mining: A survey", In: Geographic Data Mining and Knowledge Discovery. CRC Press 2001, pp. 188-217.
- [11] R. Laurini, F. Milleret-Raffort and K. Lopez, "A Primer of Geographic Databases Based on Chorems", Springer Verlag LNCS 4278, 2006, pp. 1693-1702.
- [12] W. Tobler, "Les interaction spatiale: solution de W.Tobler, Espace Populations Sociétés", 1991, pp 467-485.
- [13] R. Yann and C. Zanin Tobelem, "L'Europe dans la régionalisation de l'espace mondiale : étude des flux commerciaux par un modèle d'interaction spatiale", Géocarrefour, 2009, pp 137-149.
- [14] I. Cherni, S. Ouerteni, S. Faiz, S. Servigne, R. Laurini, "Chorems: A New Tool for Territorial Intelligence", 29th Urban Data Management Symposium, C. Ellul, S. Zlatanova, M. Rumor, Eds. London, Taylor&Francis, 2013, pp 67-76.
- [15] H. Escaith, "Statistiques du commerce international", Organisation mondiale du commerce, 2012.

The Knowledge Reuse in an Industrial Scenario: A Case Study

Gianfranco E. Modoni

Institute of Industrial Technologies
and Automation
National Research Council
Bari, Italy
gianfranco.modoni@itia.cnr.it

Enrico G. Caldarola

Institute of Industrial Technologies
and Automation
National Research Council
Bari, Italy
Department of Electrical Engineering
and Information Technologies
University of Naples "Federico II",
Napoli, Italy,
enrico.caldarola@itia.cnr.it

Walter Terkaj, Marco Sacco

Institute of Industrial Technologies
and Automation
National Research Council
Milano, Italy
{walter.terkaj,
marco.sacco}@itia.cnr.it

Abstract—Many recent research works have investigated the potential of an Ontology-based approach to support the standardization of the information in industrial scenarios. A key success factor in this regard is the effective and efficient reuse of existing knowledge sources because the building of new ones from scratch is an expensive and time-consuming activity. Although there are many advantages for reuse in the knowledge engineering, the topic is not explored in depth and the current state of the art in this field demands further investigation. The study introduced in this paper addresses the applicability of an approach to knowledge reuse based on the combination of existing techniques and methods proposed in the literature. Specifically, the experimentation has been carried out in the context of the ongoing European research project Apps4aME, where an automated framework for knowledge reuse has been tested and validated, focusing in particular on the food knowledge domain. The paper summarizes the main results of the research work and includes the emerging issues as well as some proposals to overcome them.

Keywords—*knowledge reuse; ontology engineering; semantic matching.*

I. INTRODUCTION

Reuse is an intrinsic practice in traditional engineering fields. The designers of different disciplines, from mechanics to electronics, apply it successfully whenever they build a new component, thus saving cost and time and improving the overall system quality. In this regard, an important example in the context of electronics is the reuse of standard components with well-documented and well-defined interfaces during the design of electrical circuits. Contrary to the traditional disciplines of engineering and despite intense efforts, reuse remains an underexplored and not standard process of the knowledge engineering. In this area, the key resources that can be reused are the reference models, which represent an abstract framework to understand significant relationships between defined concepts related to a specific domain, by developing consistent specifications.

Nowadays, a large number of reference models, covering a wide range of domains, are available in literature and can be considered a valid starting point for the knowledge reuse. In particular, at this stage, it is essential to refer to the state-of-the-art technical standards covering different domains,

e.g., the Industry Foundation Classes (IFC), the Standard for the Exchange of Product model data (STEP), and the International Society of Automation standards (e.g., ISA-95). In fact, they contribute to enable a comprehensive conceptualization of the represented domains, thus simplifying the communication and collaboration between the involved actors. The ability to perform effectively and efficiently knowledge reuse plays also a crucial role in the development of ontologies, which are one of the most debated topics in data modelling research community because they represent a potential solution to the problem of standardization of information [1][2]. In the context of ontology engineering, reuse of existing reference models has several advantages. First, it reduces the cost and the time required for the conceptualization of specific domains from scratch [3]. Moreover, it increases the quality of newly implemented ontologies as the reused components have already been validated. Finally, it avoids the confusion and the inconsistencies that may be generated from multiple representations of the same domain; thus, it strengthens the orchestration and harmonization of knowledge [4].

A lot of efforts has been devoted towards the application of knowledge reuse in the field of ontology engineering. In this regard, Pinto and Martins [5] have analyzed the process from a methodological point of view, thus introducing an approach that comprises several phases and activities. Moreover, the European research project NeOn proposed a novel methodology for building ontology which emphasizes the role of existing ontological and non-ontological resources for the knowledge reuse [6]. However, some open issues still remain, especially with regards to the difficulty of dealing with the extreme formalisms heterogeneity of the increasing number of models available in literature [3]. The absence of an automatic framework for the rigorous evaluation of the knowledge sources is also a severe limitation to overcome. Finally, some sub-processes of the reuse process have been defined and formalized only at the theoretical level. Therefore, it is essential to carry out the experimentation within practical cases.

This paper presents an extended feasibility study of an approach to knowledge reuse based on the combination of existing techniques and methods proposed in the literature. The approach has been tested on the real industrial case of the Romanian company CarmOlimp, where existing

knowledge sources have been explored for the development of new ontologies in the Food domain, thanks to the contribution of domain experts. Through the analysis of a real case study based on empirical evidences, the idea behind this study is to encourage and support the reuse of existing reference models, enhancing its real-world usability and identifying the challenges to make this practice a viable alternative to the development of ontologies from scratch.

The reminder of the paper is structured as follows. Section II describes the CarmOlimp case study, whereas Section III introduces and illustrates the knowledge reuse framework. Finally, Section IV draws the conclusions, summarizing the major findings and the future steps.

II. THE CARMOLIMP CASE STUDY

Modern enterprises have to face the challenge of handling a large amount of data, which are expressed in different and heterogeneous formats and are also distributed in various sources, while aiming at improving the effectiveness and the quality of their production processes [7][8]. This problem is relevant also for the Romanian company CarmOlimp that plays the meat market covering a large part of the meat production chain with a wide range of products (e.g., fresh meat, processed meat, dairy products, etc.). Since the effects of globalization are forcing this company to adapt its market to lower prices and high quality, it is needed to optimize its planning and monitoring activities to reduce the time for distribution, improve its packaging and optimize the monitoring of the meat temperature. The performance of such business process can be improved if they are supported by interoperable software tools. Semantic web technologies can be adopted to develop interoperable approaches [9] supporting the collaboration between all the involved actors and resources, while taking in consideration the storage of data [10], data definition via proper meta-models [11] and the inference of new knowledge [12].

The first step towards the realization of an ontology-based approach consists in the creation of an ontology, which is a common shared representation of the objects and their relationships and is intended to be used by the apps involved in the scenarios. The problem of developing comprehensive data models for various domains has already been addressed by researchers and a large number of them is available in literature. In the belief that their reuse could greatly reduce the costs of a new implementation, the next section introduces a framework that aims at identifying relevant data models for the formal conceptualization of a generic industrial case, while using CarmOlimp as a reference case study.

III. THE FRAMEWORK

The proposed framework requires as input a conceptual representation of the data model that highlights the hierarchy of concepts and their logical relations together with useful metadata description (target model). This representation can be the result of several interviews with the company's stakeholders and can be expressed in several

languages, e.g., plain-text, XSD (XML Schema Definition), UML (Unified Modelling Language) Class Diagrams, E-R (Entity-Relationship) diagrams, etc.

As shown in Figure 1, the framework requires the contribution of two different figures: domain experts and knowledge engineers. The domain experts are people who have deep knowledge of a specific domain, whereas the knowledge engineers are employed to elicit and translate this knowledge in terms of ontology axioms. The contribution of both occur during the whole process and comprises three phases, which are briefly summarized as follows:

- identification of the knowledge domains covering the target model;
- search for the candidate reference models related to each specific domain;
- selection of the proper models to be reused.

Each of these phases is described in the next subsections, showing also how it can be applied to the specific case of CarmOlimp.

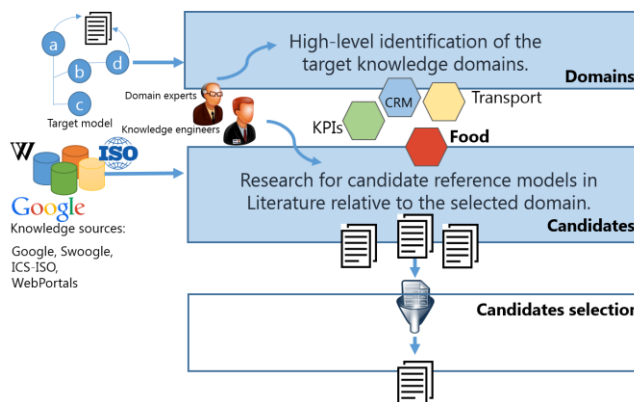


Figure 1. Framework workflow

A. The first phase: the identification of the knowledge domains

The first phase identifies the knowledge domains covering the target data model by a deep understanding of the concepts and of the metadata description within the case study. The contribution of domain experts is essential in order to clarify the meaning of some poorly defined concepts and to enable knowledge engineers to identify the right resources for this phase. Many of them are available in literature and can be used by knowledge engineers as a valuable aid to carry out the domains identification. The most important are the following:

1. Wordnet [13], a freely and publicly available large lexical database of English words.
2. General purpose or content-specific encyclopedia, e.g., Wikipedia and The Oxford Encyclopedia of Food and Drink in America.
3. Web directories, e.g., DMOZ (from directory.mozilla.org) and Yahoo! Directory.

4. Standard classifications, e.g., the International Classification for Standards (ICS) compiled by ISO (International Standardization Organization).
5. Other electronic and hard-copy knowledge sources, including technical manuals, reports and any other documentation that the domain experts may consider useful to identify the knowledge domains.

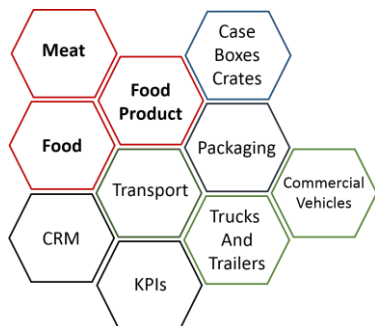


Figure 2. CarmOlimp knowledge domains

For the CarmOlimp case study, the knowledge engineers have identified the domains reported in Figure 2, which can be also mapped onto the following categories and sub-categories of ICS:

01: Generalities. Terminology. Standardization. Documentation:

- 01.040.03: Services. Company organization, management and quality;

03: Services. Company organization, management and quality. Administration. Transport. Sociology:

- 03.220: Transport
- 03.220.01: Transport in general;
- 03.220.20: Road transport;

17: Metrology and measurement. Physical phenomena:

- 17.200: Thermodynamics and temperature measurements;
- 17.200.20: Temperature-measuring instruments;

25: Manufacturing engineering:

- 25.040: Industrial automation systems
- 25.040.01: Industrial automation systems in general
- Key performance indicators (KPIs) for manufacturing operations management;

43: Road vehicles engineering:

- 43.080: Commercial vehicles;
- 43.080.10: Trucks and trailers;

53: Materials handling equipment:

- 53.060: Industrial trucks;
- 53.080: Storage equipment;

55: Packaging and distribution of goods:

- 55.020: Packaging and distribution of goods in general;
- 55.040: Packaging materials and accessories;
- 55.160: Cases. Boxes. Crates;
- 55.180: Freight distribution of goods;
- 55.180.20: General purpose pallets;

67: Food technology:

- 67.020: Processes in the food industry;
- 67.040: Food products in general;
- 67.120: Meat, meat products and other animal produce;

Since several domains have emerged during the analysis, the rest of the discussion will briefly overview their description and will focus mainly only on the food domain. This latter can be considered particularly significant for the analyzed scenario, as the core business of Carmolimp is meat processing.

B. Second phase: the search for candidate reference models

During this phase, knowledge engineers and domain experts continue to collaborate in order to identify the best tools and resources supporting the search for the relevant sources covering the selected knowledge domains. Within this study, the following sources were explored:

1. Specialized portals and websites within public or private organizations.

2. Search engines (e.g., Google, Bing, etc.), directory-based engines, e.g., Yahoo!, BOTW (Best of the Web Directory), DMOZ, etc., specialized semantic-based engines, e.g., Yummly (specialized on food), True Knowledge, etc.

3. Ontology repositories including: BioPortal, Cupboard, Schemapedia, Knoodl, etc., and search engines for semantic web ontologies, e.g., Swoogle and the Watson Semantic search engine.

4. Available standards and non-standard reference models that provide requirements, specifications, guidelines and characteristics of a service or a product (ISO standards, the IFC Industry Foundation Classes, Ansi/ISA-95, STEP, mentioned above, and the Core Product Model).

On the basis of these sources, the search has yielded a set of ten relevant candidates (Table II) for the food domain in the Carmolimp case study. They are the input for the next step of the framework.

C. Third phase: candidates selection

This phase identifies the best candidates for the formal conceptualization of the target model. The identification is performed by an initial qualitative analysis that yields the ranking of candidates according to some preference criteria. Afterwards, a linguistic analysis is carried out in order to identify the candidates which are conceptually closer to the target model (relatedness), by a measure of their similarity level.

Reference models can be rated according to relevant technical characteristics and other general information. The most important ones, considered in this study for the qualitative analysis of the sources, are shown in Table I and summarized in the following list:

(c1) **Model formality level.** It describes the formality of the conceptual model representation that can range from plain text to description logic-based languages.

(c2) **Model type generality.** It evaluates the model type from the viewpoint of its generality (upper-level model or application specific model).

(c3) **Model type structure.** It evaluates the model type from the viewpoint of its structure (simple classifications or taxonomies versus semantic enriched ontologies).

(c4) **Model language.** It describes the language used to represent the conceptual model, including RDF/OWL (Resource Description Framework/Ontology Web Language), graphic-based languages and pure text.

(c5) **Model provenance.** It evaluates the model from the viewpoint of its origin, thus giving higher rates to standards

or conceptual models authored by influential scientific groups.

(c6) **Model license.** It evaluates the availability of the conceptual model (open data-model versus proprietary and licensed models).

A higher rate will be given to formal models because the aim of the framework is to reuse existing models for a formal conceptualization of the target model.

TABLE I. CRITERIA FOR THE QUALITATIVE ANALYSIS

Criteria	Values					
Model formality level (c₁)	Formal model (first order logics-based)		Semi-formal model (RDF, XSD, graphics-based)		Informal model (text-based)	
Model type generality (c₂)	Upper level model			Domain model		Application specific
Model type structure (c₃)	Ontology		Taxonomy	Glossary	Classification	
Model language (c₄)	OWL	RDF	UML	E-R diag.	XSD	Text
Model provenance (c₅)	Public or private standardization organizations			Non-stand. research groups		Private companies
Model licence (c₆)	Open license			Proprietary		

TABLE II. CANDIDATES FOR THE QUALITATIVE ANALYSIS

Reference model	Source	Formality (C ₁)	Generality (C ₂)	Structure (C ₃)	Language (C ₄)	Provenance (C ₅)	License (C ₆)
National Cancer Institute Thesaurus Food product ontology	[21]	Formal	Domain	Ontology	OWL	Non-stand. Research	Open
AGROVOC	[22]	Semi-formal	Domain	Ontology	RDF	Non-stand. Research	Open
Linked Recipe Schema	[23]	Semi-formal	Domain	Ontology	RDF	Other	Open
BBC Food Ontology	[24]	Semi-formal	Domain	Ontology	RDF	Other	Open
LIRMM	[25]	Semi-formal	Domain	Ontology	RDF	Other	Open
The Product Types Ontology	[26]	Semi-formal	Application	Ontology	RDF	Non-stand. Research	Open
oregonstate.edu Food Glossary	[27]	Informal	Application	Glossary	Text	Other	Open
Eurocode 2 Food Coding System	[28]	Informal	Domain	Classification	Text	Non-stand. Research	Open
WAND Food and Beverage Taxonomy	[29]	Semi-formal	Domain	Taxonomy	Text	Private companies	Proprietary
Food technology ISO Standard	[30]	Semi-formal	Domain	Taxonomy	Text	Stand. Organiz.	Proprietary

With regards to the type generality, a domain model is more appropriate than upper level ontologies and application specific ontologies, because the latter may be too specific or too generic for the purpose of this study. The preferred model structure is Ontology because it generally provides logical links between concepts, thus adding more semantics to the data model than simple taxonomies and other types of classifications. The preferred representational language is RDF [14] and OWL [15] because the target model will be realized in those languages. With regard to the provenance, models delivered by standardization organizations or well-known research groups will be preferred to those produced by non-standard or unknown organizations. Finally, only open data models are passed onto the linguistic analysis, because their free availability is a mandatory requirement in this evaluation. Table I reports the values corresponding to the preference criteria mentioned above (with rates decreasing from left to right). On the basis of the criteria

previously defined, a comparison of the set of pre-selected models is carried out. In order to facilitate the analysis of the technical characteristics, a synopsis of the main outcomes of this study is reported in Table II. In view of these results, the knowledge engineer has selected the following candidate to be promoted for the linguistic analysis: National Cancer Institute Thesaurus, AGROVOC Ontology, BBC Food Ontology, Linked Recipe Schema and LIRMM Food Ontology.

The linguistic analysis uses WordNet and some of its APIs (Application Program Interfaces) [16] for estimating the linguistic matching measures between the target model and the candidate models. In WordNet, nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a different concept [17]. The synsets are interlinked by conceptual-semantic and lexical relations, thus realizing a graph-based structure where synsets are nodes and lexical-relations are the edges.

As shown in Figure 3, the linguistic analysis comprises three different steps. First, a list of terms is extracted from all the concepts of the candidate models and the target model. This process has been automatized using specific tools, e.g. XMI (XML Metadata Interchange) parsers, Ontology APIs including Jena, or Simple Access XML APIs, depending on the format used to represent the model. Then, all terms of the target model are compared to all those of each candidate model, in order to estimate the maximum Wu-Palmer [18] similarity between their synsets. This measure is calculated exploiting the Wordnet graph-based representation and indicates how much two terms are close to each other by

counting the number of edges between them and also by taking into account their proximity to the root concept of the hierarchy [19]. According to Lin [20], Wu-Palmer similarity has the advantage of being simple to calculate, in addition to its performances while remaining as expressive as the others. Table III shows an example of these calculations performed for the National Cancer Institute Thesaurus and the CarmOlimp target model. Finally, after the Wu-Palmer measures are calculated for each pair of terms, the average similarity (i.e., the sum of the Wu-Palmer similarities divided by the number of analyzed pairs) has been estimated for each candidate.

TABLE III. EXAMPLE OF WU-PALMER MEASURES

		Target Model							
		Resource	Product	Meat	Beef	Salami	Trip	Client	...
National Cancer Institute Thesaurus	Food	0.625	0.706	0.923	0.857	0.800	0.533	0.375	...
	Product	0.571	1.000	0.706	0.533	0.500	0.632	0.632	...
	Meat	0.625	0.706	1.000	0.933	0.875	0.533	0.556	...
	Lamb	0.267	0.556	0.933	0.875	0.824	0.476	0.800	...
	Poultry	0.250	0.500	0.875	0.824	0.778	0.417	0.609	...
	Chocolate	0.533	0.571	0.857	0.800	0.750	0.400	0.353	...
	Drink	0.571	0.571	0.571	0.571	0.500	0.545	0.375	...

TABLE IV. MEASURES CORRESPONDING TO THE SELECTED REFERENCE MODELS

Reference model	100% Matching concepts	Wu-Palmer > 50% Concepts	Wu-Palmer total average
National Cancer Institute Thesaurus	3	304	0.671
AGROVC Ontology	4	138	0.68
BBC Food Ontology	0	201	0.64
LinkedRecipe	0	227	0.67
LIRMM food Ontology	2	92	0.68

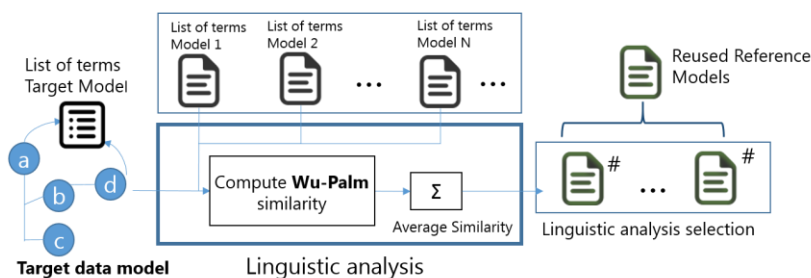


Figure 3. Selection process of the reference models

These measures are reported in Table IV together with the number of 100%-matching concepts and the number of concepts having a value of the Wu-Palmer measure greater than 0.5. The values of the average similarity are all very close and they range from 0.64 to 0.68. However, the National Cancer Institute Thesaurus proves to be the model more semantically related to the target model, since it has the largest number of 100%-matching concepts and the higher

number of concepts having a value of the Wu-Palmer measure greater than 0.5.

A preliminary validation of these results has been performed by the domain experts through an empirical method based on their expertise. Moreover, this validation has been paired with an evaluation of the accuracy of the similarity scores with respect to some existing gold standards. This verification is currently under study.

IV. CONCLUSION AND FUTURE WORKS

This work has demonstrated that an approach based on linguistic matching can help to identify the most relevant reference models that are available to cover concepts of the considered domain. Nonetheless, this approach still requires a significant amount of manual work, even when it deals with common and formal models. This requirement may be a severe limitation for a widespread adoption of the knowledge reuse, but it represents also a relevant technological gap to be addressed in future works by developing new methods and tools. Moreover, new similarity measures can be studied to improve the accuracy of the herein presented matching framework also with respect to some existing gold standards.

ACKNOWLEDGMENT

The research reported in this paper has been funded by the European Union 7th FP (FP7/2007–2013) under the grant agreement No: 314156, Engineering Apps for advanced Manufacturing Engineering (Apps4aME). The authors would like to thank CarmOlimp S.R.L. (Romania) for kindly providing information for representing the test case.

REFERENCES

- [1] A. M. Rinaldi, "A content-based approach for document representation and retrieval", In Proceedings of the eighth ACM symposium on Document engineering, 2008, pp. 106-109. Doi: 10.1145/1410140.1410163.
- [2] A. M. Rinaldi, "A multimedia ontology model based on linguistic properties and audio-visual features", Information Sciences, vol. 277, September. 2014, pp. 234-246.
- [3] E. P. Bontas, M. Mochol, and R. Tolksdorf, "Case Studies on Ontology Reuse", Proceedings of the IKNOW5 International Conference on Knowledge Management, vol. 74, June. 2005.
- [4] H. Alani, "Position Paper: Ontology Construction from Online Ontologies", Proceedings of the 15th international conference on World Wide Web, 2006, pp. 491-495. ISBN:1-59593-323-9. Doi:10.1145/1135777.1135849
- [5] H. S. Pinto and J. P. Martins, "Ontologies: How can They be Built?", Knowledge and Information Systems, Springer-Verlag, vol. 6, Issue 4, July. 2004, pp. 441-464.
- [6] M. C. Suárez-Figueroa, A. Gómez-Pérez and E. Motta, "Ontology Engineering in a Networked World", Springer-Verlag, 2012, pp. 9-34. Doi: 10.1007/978-3-642-24794-1.
- [7] G. E. Modoni, M. Sacco, and W. Terkaj, "A Semantic Framework For Graph-Based Enterprise Search", Applied Computer Science, Vol. 10, no. 4, 2014, pp. 66–74.
- [8] E. G. Caldarola, M. Sacco, and W. Terkaj, "Big Data: The Current Wave Front of the Tsunami", Applied Computer Science, Vol. 10, no. 4, 2014, pp. 7–17.
- [9] B. Kádár, W. Terkaj, and M. Sacco, "Semantic Virtual Factory supporting interoperable modelling and evaluation of production systems". CIRP Annals Manufacturing Technology, vol. 62, 2013, pp. 443-446, doi:10.1016/j.cirp.2013.03.045.
- [10] G.E. Modoni, M. Sacco, and W. Terkaj, "A survey of RDF store solutions". Proceedings of 2014 International ICE Conference on Engineering, Technology and Innovation (ICE), Bergamo, Italy, June. 2014, pp. 1-7.
- [11] W. Terkaj and M. Urgo, "Ontology-based modeling of production systems for design and performance evaluation". Proceedings of 12th IEEE International Conference on Industrial Informatics (INDIN), July 2014, pp. 748-753.
- [12] K. Efthymiou, K. Sipsas, D. Mourtzis, and G. Chryssolouris, "On knowledge reuse for manufacturing systems design and planning: A semantic technology approach", CIRP Journal of Manufacturing Science and Technology, Volume 8, January 2015, pp. 1-11.
- [13] Wordnet: A lexical database for english [Online]. Available from: <http://wordnet.princeton.edu/wordnet/> [retrieved: Jan, 2015].
- [14] R. Cyganiak, D. Wood, and M. Lanthaler, "RDF 1.1 Concepts and Abstract Syntax", W3C Recommendation, 25 February 2014. [Online]. Available from: <http://www.w3.org/TR/rdf11-concepts/>. [retrieved: Jan, 2015].
- [15] P. Hitzleret, M. Krötzsch, B. Parsia, and S. Rudolph, "OWL 2 Web Ontology Language Primer (Second Edition)", W3C Recommendation 11 December 2012. [Online]. Available from: <http://www.w3.org/TR/2012/REC-owl2-primer-20121211/>. [retrieved: Jan, 2015].
- [16] H. Shima, WS4j - WordNet Similarity for Java. [Online]. Available from: <https://code.google.com/p/ws4j/>. [retrieved: Jan, 2015].
- [17] C. Fellbaum, "WordNet: An Electronic Lexical Database, Language, Speech, and Communication Series. The MIT Press, Cambridge MA, 1998, pp. 23-61.
- [18] Z. Wu and M. Palmer, "Verb Semantics and Lexical Selection", 32nd Annual Meetings of the Associations for Computational Linguistics, 1994, pp. 133-138.
- [19] J. Euzenat and P. Shvaiko, "Ontology Matching", Springer, Heidelberg, 2007, pp. 101-103.
- [20] D. Lin, "An Information-Theoretic Definition of similarity". In Proceedings of the fifteenth International Conference on Machine Learning (ICML'98). Morgan-Kaufmann: Madison, WI, 1998, pp. 296-304.
- [21] National Cancer Institute Thesaurus [Online]. Available from <http://bioportal.bioontology.org/>. [retrieved: Jan, 2015].
- [22] Agricultural Information Management Standards AGROVOC Ontology [Online]. Available from <http://aims.fao.org/standards/agrovoc/functionalities/search>. [retrieved: Jan, 2015].
- [23] Linked Recipe RDF Schema [Online]. Available from <http://linkedrecipes.org/schema>. [retrieved: Jan, 2015].
- [24] BBC Food Ontology [Online]. Available from <http://www.bbc.co.uk/ontologie>. [retrieved: Jan, 2015].
- [25] LIRMM Food Ontology [Online]. Available from http://lov.okfn.org/dataset/lov/details/vocabulary_food.html. [retrieved: Jan, 2015].
- [26] The Product Types Ontology [Online]. Available from <http://www.productontology.org>. [retrieved: Jan, 2015].
- [27] Oregon State University Food Glossary [Online]. Available from <http://food.oregonstate.edu/>. [retrieved: Jan, 2015].
- [28] Euzenat J. Main Food Groups: classification, categories and policies [Online]. Available from <http://www.ianunwin.demon.co.uk/eurocode/docm/ec99/ecmgintr.htm>. [retrieved: Jan, 2015].
- [29] WAND Food and Beverage Taxonomy [Online]. Available from <http://www.wandinc.com/wand-food-and-beverage-taxonomy.aspx>. [retrieved: Jan, 2015].
- [30] ISO Food technology standards [Online]. Available from http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_ics_browse.htm?ICS1=67. [retrieved: Jan, 2015].

ARPPA: Mining Professional Profiles from LinkedIn Using Association Rules

Paula R. C. Silva, Wladimir C. Brandão

Department of Computer Science

Pontifícia Universidade Católica de Minas Gerais

Belo Horizonte, Brazil

paula.raissa@sga.pucminas.br, wladimir@pucminas.br

Abstract—Human resources managers design and develop professional profiles to maximize the organization workforce. These organizations typically maintain extensive static resume databases from where managers extract and analyze professional data, discovering people with appropriate knowledge, skills and experience to fulfill organizational positions. Nowadays, online professional networks, such as LinkedIn, provide a rich, dynamic, and massive scale resume database useful for professional profile analyses. Considering such massive scale databases, while manual analysis is an exhaustive and often prohibitive task, the use of data mining techniques allows managers to effectively the huge volume of data with a lower cost. In particular, for educational institutions focused on the development of persons with knowledge and skills required by organizations, the use of data mining techniques over professional networks is paramount to plan, direct and implement academic activities and curricula. In this article, we introduce ARPPA, a novel approach to discover professional profile patterns from LinkedIn by using association rules mining. Particularly, our approach crawls resumes from LinkedIn uses a multidimensional data model suitable for professional profile analyses to create and load the crawled data to a data warehouse, and extracts relevant patterns from the data warehouse using an Apriori algorithm. Additionally, we evaluate our approach attesting its usefulness to plan, direct and implement academic activities and curricula in educational institutions.

Keywords—Data-mining; association rules; LinkedIn.

I. INTRODUCTION

There has recently been a proliferation of social networks addressing the information needs of different groups of users with multiple interests, such as relationship, jobs and business [1][2]. In particular, LinkedIn stands out as the largest and most popular professional social network in the world, with more than 250 thousand of users distributed in more than 200 countries [3]. LinkedIn provides a plethora of services, such as the storage of online versions of multi-language users resume, publishing and searching job opportunities, career planning, and partnership recommendation [4].

This rich, dynamic, and massive scale source of professional information has replaced static resume databases being widely exploited by human resources managers to discover people with appropriate knowledge, skills and experience to fulfill organizational positions, and to design and develop professional profiles to maximize the

organization workforce. Particularly, for educational organizations, the use of such source of professional information to discover professional profile patterns of their current and former students is paramount to plan, direct and implement academic activities and curricula in order to meet industry requirements. For instance, analyzing professional profile from LinkedIn can help universities to invest in new educational or research lines, such as the creation of new classes or research groups for specific subjects based on trend of the former students' career. However, due to the high cost of manual data crawling, many universities do not maintain a database with updated information on professional profiles, particularly for their former students. Therefore, mechanisms to automatically crawl professional information from LinkedIn is useful for organizations interested in low cost and effective solutions to build and maintain professional profiles databases.

Crawling and mining professional profiles from LinkedIn are challenging problems. Particularly for crawling, there are duplicated data, spam, access restrictions and ambiguity issues that must be overcome. In addition, the massive scale nature of LinkedIn imposes limitations for data extraction, transformation and storage. Moreover, manual analysis is an exhaustive and prohibitive task often demanding the use of data mining techniques for fast, less expensive and more effective analytical processing. In this article, we introduce ARPPA, a novel approach to discover professional profile patterns from LinkedIn by using association rules mining. ARPPA is an acronym for “Association Rules for Professional Profile Analysis”. Our proposed approach addresses the professional profile mining problem by providing a multidimensional data model suitable for professional profile analysis and using an *Apriori* algorithm to recognize mutual implications among professional events, lastly retrieving relevant information on professional profiles. Experiments on a professional dataset crawled from LinkedIn attest the simplicity and effectiveness of ARPPA, showing that it can be used to plan, direct and implement academic activities and curricula in educational institutions.

The remainder of this article is organized as follows: In Section II, we review the related literature on data mining. In addition, we review different approaches for professional profile analysis reported in the literature. In Section III, we introduce the ARPPA approach by presenting its architecture and main components. In Section IV, we present the evaluation procedures that serve as the basis for the experiments, and we thoroughly validate our professional

profile analysis approach using a real professional dataset crawled from LinkedIn. Lastly, in Section V, we summarize our main contributions and conclusions, presenting directions for future research.

II. BACKGROUND

In this section, we review data mining from the literature. Additionally, we review the literature on data crawling and mining from social networks for professional profile analysis.

A. Data Mining

Data mining is a set of techniques to discover knowledge in a large amount of data, enabling the extraction of relevant patterns, which cannot be easily detected only by navigation or searching [5]. Typically, data mining algorithms recognize relevant patterns in datasets organized by data models suitable for effective data processing and recovery.

According to Inmon [6], the multidimensional data modeling promotes the organization of datasets using dimensions and facts to describe event occurrences, where facts are numerical measures related to events, and dimensions are properties that describe and classify the events. Multidimensional data models are used to structure and summarize datasets, presenting them in dimensional views to support online analysis. The multidimensional structure, created according to the event of interest, is commonly known as data cube. Data cubes structure data are to be viewed in multiple dimensions, in which each face of the cube represents a dimension, i.e., a perspective of an event of interest [7].

There are different approaches for multidimensional data modeling, each one based on the selection of relevant properties of the event of interest. The most used approach is the *star schema* that organizes event properties in facts and measures, linking a dimensional dataset to each fact [8]. In particular, the *star schema* is composed by one central fact table, which contains the numerical measures related to event occurrences, and a set of dimension tables [7]. A data warehouse (DW) is a n-dimensional data cube often structured using the *star schema* approach and used to support online analysis processing (OLAP) [6], creating perspective by extracting and crossing data [9]. Over DW, we can use scanning techniques to recognize data patterns and show relevant information [10].

A subset of data mining algorithms extracts relevant patterns by recognizing implication rules over data. While clustering and classification algorithms segment data into groups based on features, sequence algorithms identify frequent occurrences over the time, and association rules

algorithms use unsupervised learning techniques to identify frequent occurrences in events. The *Apriori* algorithm is an example of an association rule algorithm which extracts relevant patterns by discovering frequent association rules in events [11].

B. Crawling and Mining Social Networks

Crawling data from the Web consists in visit web servers using an automatic mechanism to collect public documents. According to Myllymaki [12], a crawler must request and store documents from web servers, extract links from the collected documents and schedule the next crawling step by using the extracted links.

Despite the inherent difficulty of crawling public data, crawling data from social networks is an even more difficult task, since the social network servers are usually not freely available for crawling [13]. Social network data may be only available for search by using an application programming interface (API), which restricts the crawling scope. For instance, the LinkedIn social network of jobs opportunities and business presents its own “People Search API” to search people, where programmers must use keywords and predefined fields, such as name, company, and school, to submit queries and receive a restricted set of professional profiles in a standardized format [1].

According to Lops [2], the proliferation of social networks generates a massive volume of data useful to learn user interests and tastes. The authors propose an approach to model user interests based on professional profiles extracted from LinkedIn. The proposed approach uses the *LinkedIn API* to get professional profiles freely available by LinkedIn users and has been used to recommend scientific articles to researchers.

In the same line, Hodigere [14] proposes an approach to predict employee careers using professional profiles extracted from a private social network. Data mining algorithms use multiple dimensions, such as positions, courses and schools, as features to rank employees by their potential of professional development.

There are different approaches reported in literature that have been using data mining techniques to discover knowledge from professional social network [15][16][17]. The Pizzato [18] applied data mining, machine learning and social network analysis on raw data extracted from LinkedIn to a people recommender system. The *SimCareers* framework [19] models member similarity over professional networks. The framework use raw data collected from LinkedIn and data mining algorithms to model and compare a sequence of work experiences finding similarities between the professional profiles.

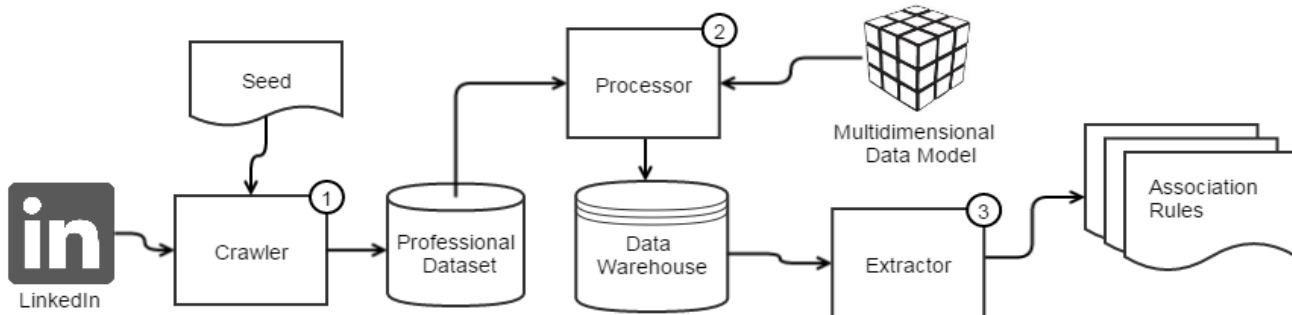


Figure1.The ARPPA architecture and main components.

III. THE ARPPA APPROACH

In this article, we introduce ARPPA, a novel and effective data mining approach to support professional profile analysis. Figure 1 presents the ARPPA architecture, including its main components.

A. Crawling LinkedIn

Step 1 in Figure 1 presents the Crawler component responsible to collect LinkedIn data. It is a Hypertext Preprocessor (PHP) component that uses the “People Search API” to retrieve professional profiles in JavaScript Object Notation (JSON) format. The crawling process is divided in two phases.

In the first phase, we use the “People Search API” to retrieve professional profiles, considering the predefined fields name and school, and particular keywords for each field. In preliminary crawling process, we observe that LinkedIn users do not properly use the name field. Sometimes users provide the first and last name but not the middle name, other times user provides the first and middle name but not the last name, increasing ambiguity. Thus, we use different name combinations to improve matching performance.

In the second phase, we collect professional profiles directly from the public LinkedIn profile pages to extract complementary data. Figure 2 presents the professional profile crawling flow used in the second phase.

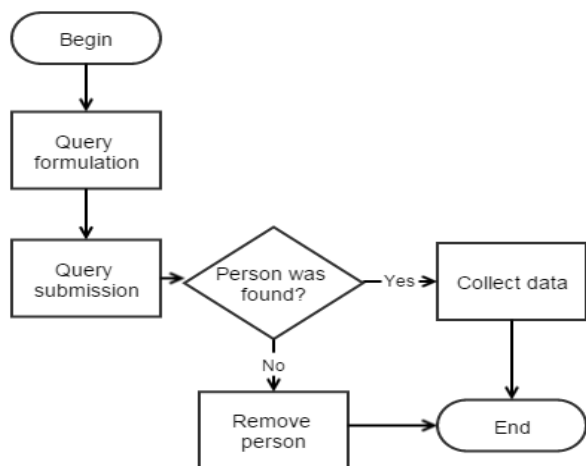


Figure 2. The professional profile crawling flow

This open crawling process is primarily necessary because users frequently disallow the access of their professional profile by the “People Search API”. In addition, we need to validate collected data in the prior phase, ensuring that the professional profiles are really related to an educational institution. We particularly process the professional profile collected in the second phase to extract the course name and the school name, comparing them with data collected in the phase one.

B. Processing Professional Profiles

Step 2 in Figure 1 presents the Processor component responsible to extract, transform and load (ETL) raw data collected from LinkedIn to a data warehouse, considering our multidimensional data model. The ETL process includes name deduplication and dimensional resolution, i.e., break one transactional entity in multiple dimensions and grouping similar instances of a dimension in one single instance. For example, the transactional entity “location” is broken into “city” and “country” dimensions, the instances of the position dimension “IT Administrator” and “Information Technology Manager” is grouped into “IT Manager”, synonyms are reduced to a single form, and typos are removed.

Figure 3 presents an excerpt of our multidimensional data model used by the Processor component to build the data warehouse.

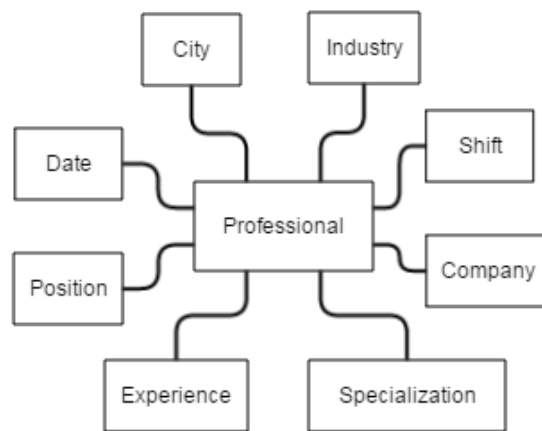


Figure 3. Multidimensional data model for mining professional profiles

From Figure 3, we observe the *star schema* used to organize professional profile data with eight dimension tables and one “Professional” fact table storing professional occurrences, suitable for online analysis processing. These multidimensional data model structures and summarizes the professional profile dataset, creating perspectives by crossing data. Table I describes each dimension from the multidimensional data model.

TABLE I. MULDIMENSIONAL DATA MODEL - DIMENSIONS

<i>Dimension</i>	<i>Description</i>
City	The city where people work.
Company	The company where people work.
Date	Semester and year of graduation.
Experience	The level related to the position dimension.
Industry	The activity area business.
Position	Job title, role.
Shift	Shift in which people attended university.
Specialization	Technology in which people are specialized.

C. *Extracting Association Rules*

Step 3 in Figure 1 presents the Extractor component responsible to discover professional patterns from the multidimensional database. In particular, we use the *Apriori* algorithm implemented in Waikato Environment for Knowledge Analysis (WEKA) to extract association rules from the data warehouse.

For each extracted rule, the *Apriori* algorithm computes metrics for analysis: confidence, conviction, leverage, and lift. The confidence measures the accuracy of the rule, i.e., the probability of occurrence of the association pattern [20]. The conviction is an alternative to confidence, also used to measure the accuracy of the rule. The leverage is a measure of divergence from the expected value [21]. Lastly, the lift, also known as interest, is used to access the standard deviation.

IV. RESULTS

To access the usefulness of the ARPPA approach for academic planning, we analyze professional patterns using a professional profile dataset collected from LinkedIn. This professional profile dataset is composed by professional data of 1,847 current and former students from the department of computer science of a major private university in Brazil. For privacy, the dataset were anonymised.

Particularly, we use the dataset as a source of information for the ARPPA approach, which extracts, transforms and loads the raw data to the DW, considering the ARPPA multidimensional data model, and uses the *Apriori* algorithm to discover association rules from DW. The association rules discovered by ARPPA were used to characterize professional behavior. Despite the discovered association rules alone do not characterize professional behavior, they point to nontrivial professional patterns that motivate further investigation.

A. *Professional Characterization*

Figure 4 presents professionals per city, considering only the top 5 cities with more professionals.

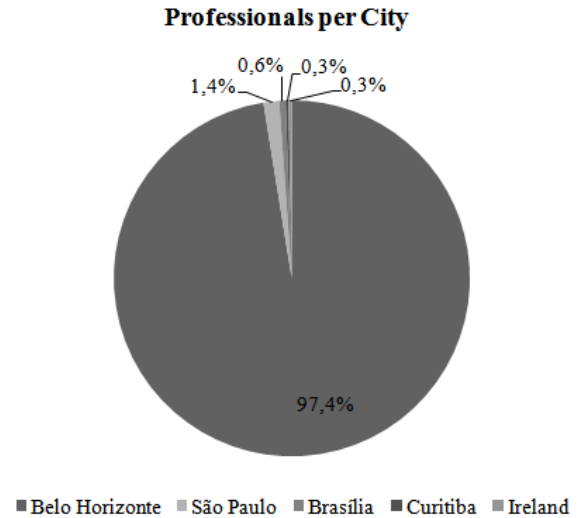


Figure 4. Professionals per city (top 5 cities)

From Figure 4, we observe that the most part of the professionals are working in *Belo Horizonte*. This is an expected behavior since the IT students sample are from the metropolitan region of *Minas Gerais*. They have studied and continue working in this region. We also observe a short percentage, but not negligible, outside Brazil. This behavior can be caused by the Brazilian government's initiatives to do partnership with abroad universities and offer scholarships in these universities. The students have to come back to finish the course in Brazil, but if they found job opportunities abroad, they commonly return to work.

Figure 5 presents the number of professional distributed by industry and shift, considering the top 5 industries.

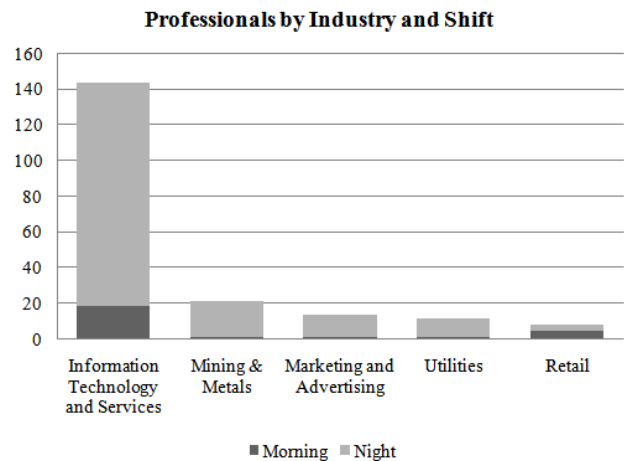


Figure 5. Number of professionals by industry and shift

Starting from the five main industries where the most part of professionals are working, we cross data with the shift that the people studied. We observe that the night shift has more professionals than the morning shift, since most

universities campus offer IT courses at night. Only one campus offers IT courses at morning. In addition, we notice that there is no difference between the industries followed by these students. The main industry where the students have actuated is “Information Technology and Services”, as expected. Moreover, there is a greater participation of professionals from the morning shift in the retail industry.

Figure 6 presents the distribution of professional distributed by positions and shift, considering the top 5 positions.

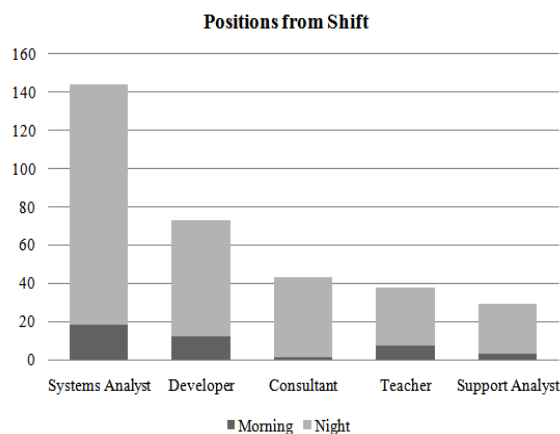


Figure 6. Number of professionals by position and shift

From Figure 6, we can see that the main position that professionals work is “Systems Analyst”. But there is a tendency that professionals follow a technical career, except for those following an academic career.

Figure 7 presents the distribution between the top five industries and the top five specializations.

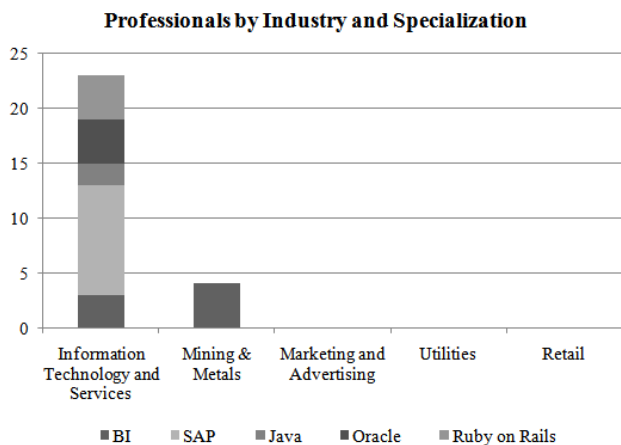


Figure 7. Relation between top 5 industries and top 5 specializations

The specialization dimension represents the technology, which professionals are expert in. From Figure 7, we notice that the main technology is “Java”, but there is a little number of professionals that is specialized in this technology. It might be caused by the many kinds of technologies that the companies have adopted. Another explanation for these values is that the computations are based on the data provide by people in their public profile;

they might not provide full information about their careers. There is a bigger concentration of technologies in industry “Information Technology and Services”. Its main cause may be the large demand for solutions to different problems; each solution is related to certain technology. The fourth and fifth industries are not related with any specialization. These industries are more specific and they demand other technologies.

Figure 8 presents the distribution of professionals by crossing between top five companies and top five positions.

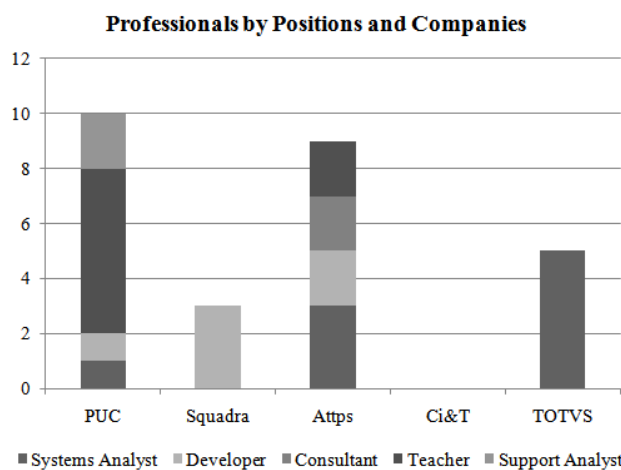


Figure 8. Relation between top 5 companies and top 5 positions

The behavior presents the main industry where the students are working: “PUC Minas” (“Pontificia Universidade Católica de Minas Gerais”), the university where they studied. The other companies are from the “Information Technology” industry. There is a little number of professionals’ concentration in each company and position. This behavior may be caused by the large distribution of the professional in many companies of different sizes. The fourth company does not hire professionals from any top five positions. The necessity for positions like “Software Engineer” and “Test Analyst” can be the main cause of this fact.

B. Professional Profile by Rules

A set of association rules show, with a minimum confidence of 0.87, which professionals following the “Systems Analyst” career are working in the “Information Technology and Services”. This is an expected behavior because “Systems Analyst” is the main position and “Information Technology and Services” is the main industry. So, we can conclude that former students are following the “Systems Analyst” career in “Information Technology” industry.

A set of association rules show, with a minimum confidence of 0.95, those students, who graduated in 2010, reached the senior level experience in Information Technology and Services industry. With this behavior we can infer that students graduated in 2010 are Senior in “Information Technology and Services” Industry.

With a minimum confidence of 1.0, an association rule shows that professional who are specialist in “SAP” technology are working in “Information Technology and Services” industry. In figure 7, we see the number of professionals who are working in each main specialization by industry, which proves the existence of this pattern of career behavior.

An association rule, with a minimum confidence of 0.97, shows that students who are in “Information Technology and Services” industry are working and living in “Belo Horizonte”. This behavior could be caused by the large proliferation of companies from this industry in “Belo Horizonte”. There is a region named “San Pedro Valley”, which has a large number of “Startups”. The development of startups has been caused by the universities and governments projects to encourage the people to undertake in this region.

V. CONCLUSIONS AND FUTURE WORK

In this article, we introduced ARPPA, a novel professional mining approach to discover professional patterns from LinkedIn by using association rules. Our approach provides a multidimensional data model suitable for professional profile analysis and uses an *Apriori* algorithm to recognize mutual implications among professional occurrences, retrieving relevant information on professional behavior. Experiments using a professional dataset extracted from LinkedIn attest the effectiveness of ARPPA showing that it is useful for educational institutions interested in planning academic activities and upgrading curricula for development of people with knowledge, skills and experience required by organizations.

Particularly, we have used ARPPA to analyze professional profiles of current and former students of the major private university in Brazil. The professional profile analyses presented in this article are samples of the possible analyses that can be performed using ARPPA. Despite the discovered association rules alone do not characterize professional behavior, they point to nontrivial professional patterns that motivate further investigation.

Lastly, considering possible directions for future research directly inspired by or stemming from the results of this article, we plan to investigate and make a comparison with other data mining algorithms to retrieve relevant information on professional behavior, such as supervised learning algorithms and neural networks. Moreover, we plan to enrich the multidimensional data model considering social data crawled from different social networks.

REFERENCES

- [1] M. A. Russell, “Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More,” O’Reilly Media, Inc, 2003.
- [2] P. Lops, M. de Gemmis, G. Semeraro, F. Narducci, and C. Musto, “Leveraging the linkedin social network data for extracting contentbased user profiles,” *ACM*, pp. 293–296, 2011.
- [3] LinkedIn, “About us,” Mar. 2014. [Online]. Available: <http://www.linkedin.com/about-us/>
- [4] D. Agarwal, “Computational advertising: the linkedin way,” *ACM*, pp.1585–1586, 2013.
- [5] R. Elmasri, *Fundamentals of database systems*. Pearson Education India, 2008.
- [6] W. H. Inmon, *Building the data warehouse*. John wiley & sons, 2005.
- [7] J. Han and M. Kamber, *Data Mining, Southeast Asia Edition: Concepts and Techniques*. Morgan kaufmann, 2006.
- [8] P. Vassiliadis and T. Sellis, “A survey of logical models for olap databases,” *ACM Sigmod Record*, vol. 28, no. 4, pp. 64–69, 1999.
- [9] S. Chaudhuri and U. Dayal, “An overview of data warehousing and olap technology,” *ACM Sigmod record*, vol. 26, no. 1, pp. 65–74, 1997.
- [10] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*, 3rd ed., ser. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2011.
- [11] R. Agrawal, T. Imieli’nski, and A. Swami, “Mining association rules between sets of items in large databases,” *ACM*, pp. 207–216, 1993.
- [12] J. Myllymaki, “Effective web data extraction with standard xml technologies,” *Computer Networks*, vol. 39, no. 5, pp. 635–644, 2002.
- [13] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, *Dbpedia: A nucleus for a web of open data*. Springer, 2007.
- [14] R. Hodigere and D. Bilimoria, “Constructing professional resource networks from career biographical data,” *IEEE Computer Society Washington, DC, USA*, pp. 1242–1247, 2012.
- [15] S. Cetintas, M. Rogati, L. Si, and Y. Fang, “Identifying similar people in professional social networks with discriminative probabilistic models,” in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’11. New York, NY, USA: ACM, 2011, pp. 1209–1210.
- [16] A. C. d. S. G. d. Santos, T. d. P. Menezes, H. R. M. da Hora et al., “Students’ and alumni’s profiles analysis through the data mining technique: a case study in the federal institute in rio de janeiro state interior,” *Research Gate*, 2014.
- [17] J. Wang, Y. Zhang, C. Posse, and A. Bhasin, “Is it time for a career switch?” in *Proceedings of the 22Nd International Conference on World Wide Web*, ser. WWW ’13. International World Wide Web Conferences Steering Committee, 2013.
- [18] L. A. Pizzato and A. Bhasin, “Beyond friendship: the art, science and applications of recommending people to people in social networks,” in *RecSys’13*, pp. 495–496, 2013.
- [19] Y. Xu, Z. Li, A. Gupta, A. Bugdayci, and A. Bhasin, “Modeling professional similarity by mining professional career trajectories,” in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’14. New York, NY, USA: ACM, pp. 1945–1954, 2014.
- [20] P.-N. Tan, V. Kumar, and J. Srivastava, “Selecting the right objective measure for association analysis,” *Information Systems*, vol. 29, no. 4, pp. 293–313, 2004.
- [21] G. Piatetski and W. Frawley, *Knowledge discovery in databases*. MIT press, 1991. G. Piatetski and W. Frawley, *Knowledge discovery in databases*. MIT press, 1991.

Towards Improving Students' Attitudes to Lectures and Getting Higher Grades –With Analyzing the Usage of Keywords in Class-Evaluation Questionnaire–

Toshiro Minami

Kyushu Institute of Information Sciences
Dazaifu, Fukuoka Japan and
Kyushu University Library, Fukuoka Japan
Email: minami@kiis.ac.jp

Yoko Ohura

Kyushu Institute of Information Sciences
Dazaifu, Fukuoka Japan
Email: ohura@kiis.ac.jp

Abstract—The eventual goal of our study is to extract useful knowledge which will help students with improving their learning performance. Towards this goal, we have studied the methods for extracting useful information about students' attitudes to the lectures they take from the lecture-related data. Such studies will contribute to capture the real status of university students, and will be able to make a very useful tool for student development in the future. In this paper, we challenge the problem of outcome/grade estimation from the text data that have been written by the students in a term-end questionnaire. First, we introduce a new concept for rating a keyword which will contribute to increasing of the grades of the students who use them. Then, we use the contribution rates and estimate the grades of students. We change the weights and compare the estimated grades with the original ones, so that we can find the optimal weights of the keywords in the proposed framework. Finally, by using the optimal rates of keywords, we compare the usage of keywords between high-graded and low-graded students.

Keywords—Text Mining; Weight of Word for Grade Estimation; Text Analysis; Educational Data Mining; Lecture Data Analysis.

I. INTRODUCTION

It has been pointed out in Japan that the students' academic skills and achievements have been decreasing. So, the universities have been putting a great amount of efforts in order to change the university professors to be able to do better lectures through the faculty development (FD) activities. However, students' academic skills do not improve accordingly.

Based on our observation, the biggest problem does not exist in the professors' teaching skills, nor in the students' academic skills and knowledge level. Main cause of this problem is rather on the students' attitudes to learning, such as eagerness to learn, curiosity to those that surround them, motivation to learning, and other mental tendencies. So, in order to pursue a solution to the problem of declining academic performances of students, it is insufficient to do efforts on FDs only. It is very important to take the matters of students (Student Development) into consideration.

Based on such recognition, our eventual goal of the study in this paper is to find out the most appropriate teaching/advising/leading methods for the students learn the most out of the lectures. In order to do this, we pursue the practical methods of extracting tips for improving lectures from data. We would help the students in learning more effectively by utilizing these tips. Our approach to this issue consists of two steps: (1) to make a student's learner model which includes attitudes to learning by proposing new concepts and measuring

indexes for them, and see what we can find, and (2) to advise the student according to his or her learner model. This approach has an advantage in terms of understandability of humans. Even though the method we are developing is a naive one, we prefer to choose the understandable method rather than applying the established and more sophisticated methods if they are less understandable for us.

As a part of such an approach, we have been analyzing the answer texts to a question, which asked the students about their looking-back evaluation of themselves and the class [10]. Such data are considered to be appropriate to analyze the students' attitudes to the lectures. In the studies so far, we have found that the students with high examination scores use the words which indicate their wide point of view. On the other hand, the students with low grades use the words closely related to the lecture. An aim of this paper is to introduce a numerical method to text analysis and confirm this finding.

In this paper, we analyze the data obtained in the lectures. Such studies of educational data analysis have been conducted in the research field of Educational Data Mining (EDM) [12]. For example, Romero et al. [13] gave a comparative study of data mining algorithms for classifying students using data from e-learning system. Its major interest is on predicting the student's outcome. Our focus is on the student's psychological tendency in learning, such as eagerness, diligence, seriousness, etc. Many studies in EDM use the target data which are obtained from learning management systems (LMSx). On the other hand, we intend to obtain our target data in everyday lectures.

Goda et al. [2] proposed a method of text analysis, where texts are given by students as the reports in the everyday lectures. Our method, on the other hand, mainly uses such data as the homework, exercise, and term-end examination, which are obtainable in ordinary lectures.

Ames et al. [1] studied in the similar motivation to ours. They investigated the students' attitudes to the class, learning, etc., based on the answers to questionnaire items. However, their underlying data were obtained by asking the students to choose the rate from 1 to 5 for each question item. In our case, even though 2 of our question items are asking to rate from 0 to 100, other questions are asking to write the students' own thought in a free-text format.

Our data analysis style is also different from the major studies in EDM. Most of them somehow intend to analyze the big data, and the data obtained automatically as log data. On

the other hand, we would rather take the approach of dealing with small data, because our target data themselves may be very small [6][10]. Also, the data we deal with are somewhat represent human students, and we, as the staff in an educational organization like university, we have to educate all of them. Thus, we have to take attention to all the data as well, even if they are located in the far-away areas from the central area, because they represent one or more students.

We have been taking such an approach in library data analysis. In previous studies [4][7][8], we took library's loan records as the target data and analyzed them by proposing new concepts, such as expertise levels of books and library patrons. From these experiences we are convinced that such an approach is useful also for other types of small data. So, we take the same approach in our lecture data analysis.

In order to achieve our aim, the rest of the paper is organized as follows. In Section II, we describe the data we use for analysis. In Section III, we present our interesting findings in our previous studies, such as [10][11]. In Section IV, we conduct the analysis by focusing attention to the words used by the students in the answer texts of a question about their final evaluation, which asked the students to evaluate their achievements of the class. We propose a process model for estimating the grades of students, and perform the process. Then, we change the parameter in order to find the optimal results. We discuss the usage of keywords in this optimal case. Finally in Section V, we conclude the discussions and findings in this paper.

II. TARGET DATA

The data used in this paper came from the class of "Information Retrieval Exercise" in 2009 in a women's junior college [5][6][10]. The total number of students who attended the class was 35. They were year 2 students and going to graduate. The most important aim of the course was to let the students become expert information searchers so that they had enough knowledge about information retrieval, search, finding, and also had enough skills in finding appropriate search engine site and search keywords based on the understanding of the aim and the background of the retrieval.

The term-end examination of the course consists of 3 questions. The first question is to ask them to find the Web sites of search engine, and to summarize their characteristic features. The second question is on finding the Web sites on e-books and on-line material services. The third question is to find and argue about the information crimes in the Internet environment. The aim of these questions is to evaluate the skills on information retrieval, including the skills for planning and summarizing. These skills are supposed to have learned and trained in the course, through their exercises in the classes and in the homework assignments. We use the score of term-end examination as the measure for the student's achievement.

We also asked the students to answer the questions as the overall evaluation of them for the course. They are: (Q1) what the student has learned in the lectures, (Q2) good points of the lectures, (Q3) bad points that need to be improved, (Q4) to score the course as a whole with the numbers from 0 to 100, where the pass level is 60 as in the same way to the examination score, ..., (Q11) to score the student herself of her outcome and her attitude toward the course from 0 to 100 as in the same way as in Q4, etc.

III. FINDINGS IN OUR PREVIOUS ANALYSIS

We will illustrate what we have done and have found in our previous studies. Our study in Section IV is carried out based on these achievements.

A. Analysis of Numerical Items

As the first example of the results obtained in the previous analysis, Fig. 1 shows the correlation between the self-evaluation scores (x-axis) and the examination scores (y-axis). The former data are obtained as the answer to Q11. Actually, the data for about 60% of students are shown in the figure, because only 21 students gave answers to the scoring questions Q4 and Q11.

It is interesting to see that the students who have higher examination scores, i.e., those who locate above the line for the average score 71, evaluates themselves mostly in the range between 40 and 80. Thus, the range size of the self-evaluation is as wide as about 40.

On the other hand, the students who have lower-than-average scores evaluate themselves relatively high. With one exception of the student who has the lowest self-evaluation score, the rest of them mostly evaluate themselves more than average score.

In other words, we can say that the students who have high examination scores evaluate themselves from a very low score up to a very high one, which means that those students who evaluate low would have the self-image that "I am the person who can do better than what I have been doing." These students have a good desire of self-improvement.

On the other hand, the students who have poor performance seem to believe in themselves without evidence, and evaluate themselves something like, "I do fairly well in my study." Another possibility is that they actually recognize very well about their poor efforts and poor performance. Still, or maybe because of it, they wanted to believe themselves strongly, that they are not very poor in their efforts, instead of admitting their poor efforts. In this way, they could avoid facing what they really were, and keep their pride. As a result of such a phenomenon, the correlation coefficient between the self-evaluation score and the examination scores becomes a negative value of -0.1 .

In our previous analysis, we have also investigated the relations of attendance and homework scores, together with the scores in the questions of Q4 and Q11. We have found that a notable number of students are just attending the lectures and

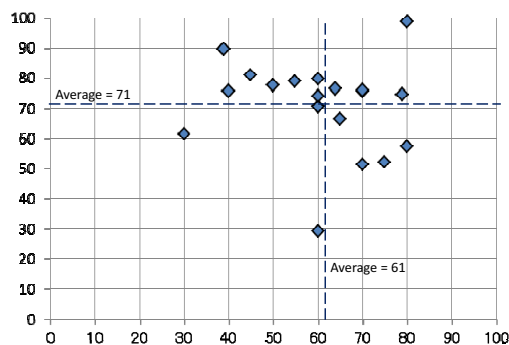


Figure 1. Correlation between self-evaluation scores (x-axis) and examination scores (y-axis).

do their homeworks. Probably because they do not intend to learn seriously, they could not learn in the real sense. Thus, quite a lot of students' efforts are rather superficial and they do not affect to the true improvement of their academic skills [9].

Furthermore, we have a result, which indicates that students concern more on their efforts than on their achievements, which is an important result for the lecturers to know. See the papers [5][6][9][10] for more findings in our analysis.

B. Analysis of the Relationship between Students' Viewpoint and Outcome from the Answers for Question 1

The first question of the term-end questionnaire was "Q1: What did you learn in this class? Did it help you?" It is the question to ask the students for summarizing what they have learned in the lectures of the course, and they express in their own words. Thus the answers to the question might express their understanding, recognition, point of view, and so on, in the course. Because the data are in free text format, it is more difficult to analyze and extract useful information out of them than the data in numerical format. However, at the same time, they are very appropriate to know about the students on their mental statuses that are normally hidden in their minds. We want to extract information such as which teaching materials attract them, in what attitudes they had in the class, how they felt by attending the class, etc.

As the first step to conduct the text data analysis, we need to transform them into the data that can be used in our analysis method. As an approach to the free text analysis, we took attention to the usage of words and phrases in the texts. The words used by the students might somewhat reflect their own views and attitudes to the course. Also, in order to obtain more subjective results, it is preferable to extract the words from the text data and analyze them than to extract the students' attitude data in a manual methods by us humans.

We chose the KH Coder [3] as the analysis tool. KH Coder is a free software equipped with the facilities of morphological analysis for Japanese language. This facility is very important in dealing with Japanese texts, because Japanese has no word divider, like the space in English. Thus word segmentation is a big issue in natural language processing. KH Coder can extract words together with doing statistical analysis including correspondence analysis. In our analysis, we took the answer text of each student as one document for KH Coder.

C. Extraction of Words which Appear in the Answers

Table I shows the words that appear in the texts more than 5 times and their number of occurrences, in the decreasing order of the number. We can see that the words related to the lectures appear in high frequencies. For example, the word "Search" appears 88 times in the answers to Q1, which is the most frequently used one among all words. Also the words "Information", "Library" and so on appear in the list. The lecture-related words are 6 (20%) among 30 words, whereas 4 (29%) among 14 words with frequencies more than 10.

D. Correspondence Analysis of Words and Students

It is important to know not only the words themselves but also their relations with others, such as between word and word, between word and student. Analysis of such association may give us more useful information about students and their attitudes to learning.

TABLE I. EXTRACTED WORDS AND THEIR OCCURRENCES (FREQ.> 5)

Word	Freq.	Word	Freq.	Word	Freq.
Search	88	Way	16	Think	8
Class	37	Examine	16	Do	8
Information	37	Keyword	13	Get	8
I think	34	Are various	11	Various	7
Library	33	Use *	10	Feel	7
Learn	32	Help	10	Function	7
Know	30	Necessary	9	Result	7
Myself	21	Use *	9	Important	7
How	21	Internet	8	Opportunity	6
Now	17	Personal Computer	8	This time	6

* Different words in Japanese

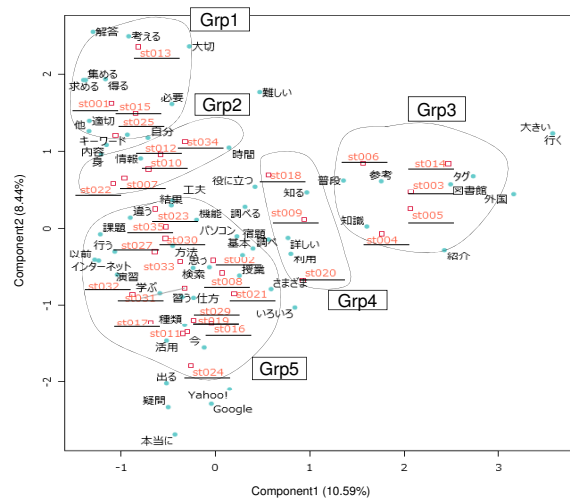


Figure 2. Correspondence analysis map of related words with students.

Fig. 2 shows the results of the correspondence analysis in a two-dimensional principle component space. The words in the figure are those occurring more than 2 times. The name in the form st0** represents a student. Then we divide the students into 5 groups manually, from Grp1 to Grp5. The total number of students who appear in the figure is 33 because two students did not answer to the question Q1. Table II shows the features of the groups, such as the numbers of students of the group, the member's average examination scores, and their variances. The numbers of the group members range from 3 (Grp4) to 16 (Grp5). For the average scores, Grp3 is the highest with the score 83.5, which locates in the upper-right part in Fig. 2. Grp5 takes the lowest average score with 59.3 which locates in the lower-left part. The P-value is 0.0469, and thus the assertion that there are differences between the averages of these examination scores of five groups is statistically significant at the 5% level [10].

E. Relation between Used Words and Examination Scores

Next, we deal with the correlations between the characteristic words that appear in the answers to Q1 and the examination scores. Table III shows the high ranked characteristic words of some of the students in the decreasing order in the Jaccard similarity measure. Note that the Jaccard similarity of student p and q is defined as the ratio of the number of words which

TABLE II. ANALYSIS OF VARIANCE TABLE OF 5 GROUPS.

Group No.	Number of Members	Average	Variance
Grp1	4	65.2	27.3
Grp2	5	70.5	98.8
Grp3	5	83.5	107.2
Grp4	3	69.8	68.7
Grp5	16	59.3	335.3

TABLE III. EXAMPLE CHARACTERISTIC WORDS OF STUDENTS IN GRP3 AND GRP5

Grp3				
st005 99	st006 76	st003 77	st004 76	st014 90
Foreign●	Library▲	Put together	Show	Go
Library▲	Individuality	Country	Limit	Foreign country
Latest	Summary	Box	Interesting	Automatic
Effort	Take	HP	Photo	Especially
World	Relationship	Closed■	Introduction	Completely
See	Plus	Books■	Familiar	Lending■
IC■	Whole country	Appear	Japan	Electronic●
Tags■	At the same time	Root	Copyright●	Large
Various	Reference■	Tackle	Every time	Usually
Feel	Also	Province	Learn	Library▲

Grp5 (5 students out of 16)				
st019 52	st030 27	st031 34	st032 29	st035 51
Old days	Result	Word	Internet●	Draw
Question	Do	Since	Use of	Home
Really	Know	Approach	I	Work
Now	Information▲	Motivation	Destination	Many
Think	Learn	Stimulus	Use	Get used to
Learn	Search▲	Not at all	Job hunting	Listen
Respond		Different	Future	Touch
Answer		Frequency	Received	Utilize
Prior		Desk	Exercises	Vaguely
Current		A lot	Previous	Schoolchildren

are commonly used by p and q against the total number of words which are used either or both of p and q. The words marked with “●”, “▲” and “■” in Table III indicate that they are classified as the general word, frequently used words that relate to the lecture, and the words closely related to the subject, respectively. The value in the right-hand side of (st0***) represents her examination score.

For Grp3 (with the highest examination score) characteristically use the technical terms and those words from the broader point of view in comparing Japan and the world such as “Foreign,” “National,” and “Japan.” It is interesting to see that the words which are relating to the homework assignments do not appear in Grp3. Thus, we can see that the students in Grp3 attended the lectures with the attitude of learning in a broad perspective.

The students in Grp5 (with the lowest examination score) use quite a lot of frequently-used general words, and do not use technical terms at all. It is interesting to see that many students use the words they have learned in the lectures; e.g., “Learn,” “Master,” “Study,” “Useful,” and “Use.” So, we can guess they took too much attentions to the words themselves and did not pay much attention to their background, their relation to the related concepts, their values in our society, etc.

IV. GRADE ESTIMATION OF STUDENTS BY THE USAGE OF WORDS IN THE QUESTIONNAIRE ANSWER

As we have shown in the previous section, there exists some amount of relationship between the grades of the term-end examination and the usage of words in the answers to Q1 [10]. For example, the students in Grp3, the group with the highest average score, used the words that indicate their interest to the world, such as “Foreign,” “Overseas,” “Japan,” and “National.” They also used more technical terms and characteristic words than other group members.

On the other hand, the students in Grp5, the one with the lowest average score, used general words and popular words, such as “Remember,” “Learn,” “Useful,” and they did not use technical terms. It is interesting to see that from the words used by the Grp5 students, they look like good students because they used the words directly related to the lectures. However, considering their poor achievements, they might have learned rather superficially in the lectures. They might thought it was very important in the lectures to attend the class regularly and remembered what the teacher talked about. They would not think it was important to have a wide view on the subjects they were supposed to learn and thought hard with their own brains. We investigated further on how much the word types affect to the students’ achievements in [11].

A. Grade Estimation of Students

Next, we estimate the students’ grades according to the framework illustrated in Fig. 3. The process consists of two steps. Step 1 is the process of assigning the weights to the keywords that appear in the texts. The weight is calculated so that it reflects the closeness of the keyword and the students who use it, and the weight becomes higher if the students took higher examination scores. The arrows in the left part of the figure illustrate this relationship.

Step 2 is the process of estimating the grades of the students. We also use the relationship between the words and the student in this step. If a student used the words which have higher weights, then the student’s estimated score becomes higher. The most typical and simple method might be to distribute the weight of a keyword to the students who have used it in the equal rate, i.e., to give the keyword weight equally to its related students by dividing the weight by the number of students who used it. Then, we calculate each student’s estimated score by summing up all the score values

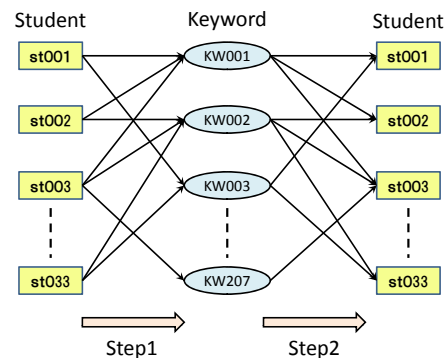


Figure 3. The method of the grade-estimation of students.

given by the words which the student used. We would like to make this simple definition more sophisticated.

Let $\{s_1, s_2, \dots, s_m\}$ and $\{w_1, w_2, \dots, w_n\}$ be the set of students and keywords, respectively, and let us define $A = (a_{i,j})$ so that $a_{i,j} = 1$ if and only if the word w_i is used in the texts provided by the student s_j , and $a_{i,j} = 0$ otherwise, for $1 \leq i \leq n$ and $1 \leq j \leq m$. For Step 1, we define a vector $S_1 = (g_1, g_2, \dots, g_m)$ so as g_i is the examination score of the student s_i . We define a matrix $B = (b_{i,j})$ by setting $b_{i,j} = a_{i,j} / \sum_{1 \leq k \leq m} a_{i,k}$, so that the grade of a student is equally distributed to its related keywords, and $b_{i,j}$ is the distributed grade value to the word w_i from the student s_j .

Using this matrix we define the vector $W_1 = (c_i)$ of estimated grade value of keywords. Let k_i be set as the sum of the distributed grade values assigned to the keyword w_j ; that is $W_1 = BS_1$, or:

$$c_i = \sum_{1 \leq k \leq j} b_{i,k} g_k \quad \text{for } 1 \leq i \leq n \quad (1)$$

Table IV shows the list of words for the weight ≥ 0.0059 in its decreasing order. The weights here are normalized so that the total sum of the weights becomes 1. By using the weights of the words, we can estimate the grade of a student by summing up the weights of the words used by the student.

$$S_{E1} = A^T W_1 \times (t/u) \quad (2)$$

where t is the sum of the original grades, and u is the sum of the components of $A^T W_1$.

Table V shows the resulting estimated grades and their original grades of students in the decreasing order of the original grades. Fig. 4 shows the correlation diagram that shows the mutual relation between the original grades and the estimated grades of students.

TABLE IV. WEIGHTS OF WORDS. (RATE ≥ 0.0059)

No	Words	Rates	No	Words	Rates	No	Words	Rates
1	Newest	0.0071	8	Especially	0.0064	15	Layout	0.0059
2	Tag	0.0071	9	Feel	0.0064	16	Master	0.0059
3	World	0.0071	10	Library	0.0063	17	Recently	0.0059
4	Overseas	0.0068	11	Tackle	0.0063	18	Report	0.0059
5	Automatic	0.0064	12	See	0.0061	19	Server	0.0059
6	Big	0.0064	13	IC	0.0059	20	Start	0.0059
7	Electronic	0.0064	14	How to write	0.0059	21	Study	0.0059

TABLE V. ESTIMATED GRADES OF STUDENTS USING THE WEIGHTS OF WORDS.

No.	Student	Original	Estimated	No.	Student	Original	Estimated
1	st005	99	94	18	st013	69	49
2	st014	90	89	19	st017	66	63
3	st002	82	84	20	st018	66	66
4	st033	81	78	21	st020	64	60
5	st010	80	81	22	st025	61	66
6	st009	79	79	23	st001	60	66
7	st011	79	82	24	st016	57	62
8	st003	77	83	25	st023	56	51
9	st008	77	64	26	st029	55	63
10	st004	76	79	27	st034	54	60
11	st006	76	82	28	st027	53	18
12	st021	76	75	29	st019	52	53
13	st007	75	74	30	st035	51	57
14	st024	75	73	31	st031	34	44
15	st022	74	76	32	st032	29	43
16	st012	71	72	33	st030	27	29
17	st015	71	74				

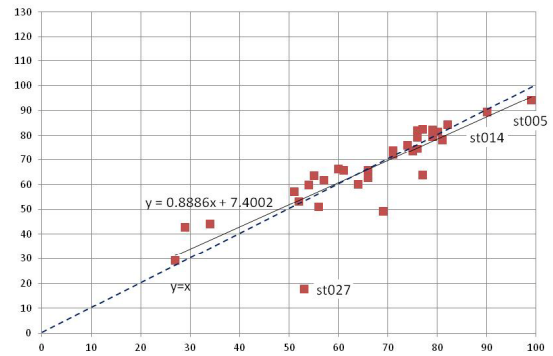


Figure 4. Correlation diagram between the original (x-axis) and the estimated (y-axis) grade.

B. Weight Adjustment for Optimal Estimation of the Students' Grades

The results obtained so far look good enough for estimating the students grades from the weights of keywords. However, the current weighting is proportional to the usage ratio of the word, and rather a simple method of weighting. Thus, it may be possible to get better result by taking different weighting method. We take an approach of introducing a parameter for changing the weighting, and find out the optimal value of the parameter for better weighting.

In order to pursue this scenario, we extend the definition of S_1 to the definition of S_n with the parameter n by: $S_n = (g_1^n, g_2^n, \dots, g_m^n)$, so that the definition of S_1 , which we have used so far, becomes a special case of S_n when $n = 1$.

Fig. 5 shows the change of correlation coefficient between the original and the estimated grades as n varies from 0 to 10. The optimal case in our framework comes when $n = 2.22$, and its correlation coefficient is 0.922.

C. Correlation between Original and Estimated Grades in the Optimal Case

Fig. 6 shows the correlation diagram of the grade-estimation of students when $n = 2.22$ and thus the correlation is optimal. Table VI shows the weights of keywords in this case. We can see that the keywords characteristically used by the students in Grp3 (with the highest average examination score) have high weight in general. For example, the word "Overseas" has the weight 0.0105, which is the 4th highest keyword weight. Other keywords "Japan" and "National" have 74th and 83rd highest among 206 keywords.



Figure 5. Change of correlation coefficient by the value of n .

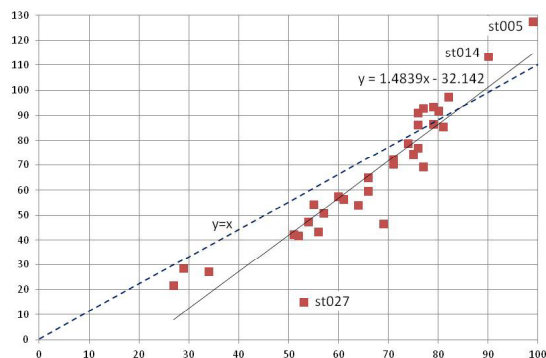


Figure 6. The optimal grade-estimation result (y-axis) represented as correlation diagram with the original grade (x-axis).

TABLE VI. WEIGHTS OF TOP MOST WORDS IN THE OPTIMAL CASE. (RATES ≥ 0.0069)

No	Words	Rates	No	Words	Rates	No	Words	Rates
1	Newest	0.0105	8	Especialy	0.0085	15	Layout	0.0069
2	Tag	0.0105	9	Feel	0.0084	16	Master	0.0069
3	World	0.0105	10	Library	0.0083	17	Recently	0.0069
4	Overseas	0.0105	11	Tackle	0.0082	18	Report	0.0069
5	Automatic	0.0105	12	See	0.0075	19	Server	0.0069
6	Big	0.0085	13	IC	0.0074	20	Start	0.0069
7	Electronic	0.0085	14	How to write	0.0069	21	Study	0.0069

For the keywords preferably used by the Grp5 students, “Remember” has the ranking order 44, followed by “Useful,” “Use,” and “Learn” with the ranking order 100, 164, and 171, respectively. Thus, from the numerical data also, it has been clearly shown that the students in Grp3 used the words which have high weights, and those in Grp5 used the words which have small weight to the grades.

V. CONCLUSION AND FUTURE WORK

We have been studying the student’s attitudes toward the lectures so far. In this paper, we conducted a study of estimating the grades of students in terms of the weights of words, which are calculated according to the frequencies of words used by students. Beginning from a simple method of linear weighting of words, we extended it to a more sophisticated method by using a parameter for adjusting the weighting. By changing the value of the parameter, we could find the value of the parameter which gives the optimal weighting of words for grade estimation. Then we discussed the words and their weights in the optimal case.

Our eventual goals in the research topic of this study are two-folds: The first one is to find new facts and tips for helping our students with more effective learning, and the other is to develop new concepts and measuring methods which can be used for the first goal. Thus, understandability is very important in our study. This is the reason why we rather choose naive methods of analysis than to use more sophisticated, but less human understandable methods.

Even though our current status of study is in a very beginning stage, the methods developed in the studies so far have shown high potential of our methods. It will become a necessary knowledge management tool for student development [7] in the near future, because it is a very important topic for the institutional research (IR) for universities [6].

We have the following study topics for the future: (1) To develop a method to devise new ideas further, and to perform refinement of dedication to the study of student effort, and attitudes to learning, especially further analysis of the text. Use of other similarity measure like pointwise mutual information (PMI) is a possible candidate. Also, it is worth comparing our model with other types of models. (2) By collecting data from a different class, to analyze them, and to verify if the results of this study are also holds. Also, it is important to find out the characteristic features of each class by comparing them. It will be interesting to investigate what features are gender-specific. (3) To generalize the analysis methods, and to integrate them into an automated data analysis system.

ACKNOWLEDGMENT

This work was supported in part by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C), 24500318, 2014.

REFERENCES

- [1] C. Ames, and J. Archer, “Achievement Goals in the Classroom: Students’ Learning Strategies and Motivation Processes,” *Journal of Educational Psychology*, Vol.80, No.3, pp.260–267, 1988.
- [2] K. Goda, S. Hirokawa, and T. Mine, “Automated Evaluation of Student Comments on Their Learning Behavior,” *12th International Conference on Advances in Web-Based Learning (ICWL 2013)*, Oct. 2013, pp. 131–140.
- [3] K. Higuchi, “KH Coder.” <http://khc.sourceforge.net/en/> [accessed: 2014-12-07]
- [4] T. Minami, and K. Baba, “Investigation of interest range and earnestness of library patrons from circulation records,” *International Conference on e-Services and Knowledge Management (ESKM 2012) in IIAI-AAI 2012, IEEE CPS, DOI 10.1109/IIAI-AAI2012.15*, Sep. 2012, pp. 25–29.
- [5] T. Minami, and Y. Ohura, “An attempt on effort-achievement analysis of lecture data for effective teaching,” *Database Theory and Application (DTA 2012)*, in T.-h. Kim et al. (Eds.): *EL/DTA/UNESST 2012, CCIS 352*, Springer-Verlag, Dec. 2012, pp. 50–57.
- [6] T. Minami, and Y. Ohura, “Towards Development of Lecture Data Analysis Method and its Application to Improvement of Teaching,” *2nd International Conference on Applied and Theoretical Information Systems Research (2ndATISR 2012)*, Dec. 2012, 14pp.
- [7] T. Minami, “Profiling of Patrons’ Interest Areas from Library’s Circulation Records – An Approach to Knowledge Management for University Students -,” *The Fifth International Conference on Information, Process, and Knowledge Management (eKNOW 2013)*, Feb. 2013, pp. 45–50.
- [8] T. Minami, “Interest Area Analysis of Person and Group Using Library’s Circulation Records,” *IADIS International Conference Information Systems (IS 2013)*, March 2013, pp. 215–222.
- [9] T. Minami, and Y. Ohura, “Lecture Data Analysis towards to Know How the Students’ Attitudes Affect to their Evaluations,” *8th International Conference on Information Technology and Applications (ICITA 2013)*, July 2013, pp. 164–169.
- [10] T. Minami, and Y. Ohura, “Investigation of Students’ Attitudes to Lectures with Tex-Analysis of Questionnaires,” *4th International Conference on E-Service and Knowledge Management (ESKM 2013)*, Sep. 2013, 7pp.
- [11] T. Minami, and Y. Ohura, “A Correlation Analysis of Student’s Attitude and Outcome of Lectures –Investigation of Keywords in Class-Evaluation Questionnaire–,” *Advanced Science and Technology Letters (ASTL)*, Vol.73 (FGCN 2014), Dec. 2014, pp. 11–16.
- [12] C. Romero, and S. Ventura, “Educational data mining: A survey from 1995 to 2005,” *Expert Systems with Applications*, Vol. 33, Issue 1, July 2007, pp. 135–146.
- [13] C. Romero, S. Ventura, P. Espejo, and C. Hervás, “Data mining algorithms to classify students,” *1st International Conference on Educational Data Mining (EDM 2008)*, June 2008, pp. 8–17.

Supporting Provenance in Climate Science Research

Brett Yasutake, Niko Simonson, Jason Woodring, Nathan Duncan, William Pfeffer,
Hazeline U. Asuncion, Munehiro Fukuda, Eric Salathe

School of Science, Technology, Engineering, and Mathematics
University of Washington Bothell
Bothell, WA, USA

{yasutake, nikouw, jman5000, njd91, wpfeffer, hazeline, mfukuda, salathe}@u.washington.edu

Abstract—While the data produced by climate models exponentially grows in size and complexity, the ability of researchers to analyze available data lags. Existing tools for climate analysis that capture provenance are generally implemented on supercomputing clusters. Provenance is often difficult for a researcher to analyze due to its sheer volume. In contrast, our Pacific Northwest Climate Analysis (PNCA) Tracker is a lightweight, provenance-aware parallel system that allows researchers from smaller facilities to quickly develop custom analysis tools while enabling them to easily verify their datasets. This technique modularizes the captured provenance, allows researchers to customize the provenance collection, and efficiently collects provenance within a parallel and distributed environment, made possible by the use of the Multi-Agent Spatial Simulation (MASS) library. It is designed to be highly extensible by minimizing dependencies within the architecture. We demonstrate that our tool is potentially accessible to a wider range of researchers and is highly efficient compared to the commonly used climate analysis tool, Network Common Data Form (NetCDF) Operators or NCO. Finally, we discuss how provenance concepts in PNCA Tracker map to the W3C PROV.

Keywords—data provenance; climate science; parallelization; big data.

I. INTRODUCTION

Regional and global climate models produce a vast array of model output data simulated for periods of years to centuries. The complexity and resolution of climate models is steadily increasing, incorporating more complex and realistic methods. Analysis of regional climate data can yield important discoveries that have wide-ranging impacts; however, climate data have grown to a scale that makes them difficult to analyze without sophisticated computational tools [13]. On top of this, small to medium-sized climate research labs are resource-constrained. These labs have access to limited computational infrastructure and their computational resources are generally focused on the analysis of data or running simulations, not on administering or maintaining specific technologies. Thus, these research labs must be able to quickly build one-off analytical tools (e.g., scripts or homegrown software), run them efficiently on large datasets, and capture useful, understandable provenance to support various analysis tasks. Data provenance is defined as the “origin or the history of data” [18]. Without data provenance, it is difficult to assess the results of a simulation.

Meanwhile, current provenance tools for climate research generally cater to labs equipped with high-end computing resources. As a result, current provenance support in this domain assumes access to High Performance Computing (HPC) resources [20] or to specific platforms (e.g., scientific workflows [11] or databases [7]). Without access to these resources or staff to operate these platforms, these climate researchers lack tools to capture provenance.

To bridge this gap, we created a novel technique called Pacific Northwest Climate Analysis (PNCA) Tracker, a platform-independent provenance tool which allows climate researchers to easily integrate provenance capture with their own analysis tools in a lightweight manner. PNCA Tracker leverages the easy-to-use Multi-Agent Spatial Simulation (MASS) library for parallelization [8] to enable researchers to rapidly develop analytical tools that can be immediately applied to a large body of climate model output data efficiently. PNCA Tracker provides simple provenance collection mechanisms for the overall framework, which runs and parallelizes researcher-implemented modules for the analysis of climate data.

The contributions of this paper are as follows: (1) an accessible, adaptable, and scalable technique to support data provenance in climate research, (2) a loosely-coupled and customizable tool framework that enables researchers to “plug-in” their custom analysis tool and off-the-shelf visualization tools, (3) a set of evaluations that suggest the viability of our technique, and (4) a discussion of provenance concepts in relation to the community standard W3C PROV.

This paper is organized as follows. The next section compares our technique to existing provenance collection techniques. It also provides background information on existing techniques used with PNCA Tracker. Section III provides a motivation for our technique. Section IV contains details about PNCA Tracker and its tool support is described in Section V. Section VI covers our set of evaluations. Section VII discusses how our provenance concepts map to the W3C PROV. The paper concludes with future work.

II. RELATED WORK AND BACKGROUND

This section discusses closely related techniques and background on techniques used with the PNCA Tracker.

A. Provenance Techniques for Climate Research

Many provenance collection techniques for climate science rely on access to HPC resources. For example, one

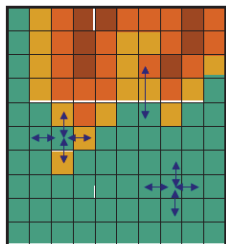


Figure 1. Visual representation of the MASS simulation space.

technique uses a high-end supercomputing cluster [20]. Milieu is a lightweight, unobtrusive provenance collection tool, but is currently implemented on the National Energy Research Scientific Computing Center's cluster [7]. Our technique does not require the availability of these types of resources.

Some techniques require the availability of specific technologies. Systems such as Milieu store data in specific databases that require native support [7]. Ultrascale Visualization Climate Data Analysis Tools (UV-CDAT) outputs application-specific VisTrails workflows [11]. These two approaches allow for very granular and highly indexed provenance that increases the value of the provenance data collected. Another technique provides provenance support to a middleware technology [15]. However, our technique does not require specific technologies to access the stored provenance, as we discuss in the next section.

Many of the software tools that support the analysis of large-scale data require some computer science background. For example, Hadoop is a well-known software package for analyzing large-scale data. HadoopProv, which supports provenance in Hadoop, also requires users to configure and monitor computing nodes in a cluster [6].

Meanwhile, SciHadoop was developed to support the array structure of data in a parallel environment [12]. Parallel NetCDF [14] and Message Passing Interface (MPI) [1] also provide parallel processing. However, these programming environments do not have built-in provenance support and, thus, must be user-customized with processor-aware code [5], which again requires non-computer scientists to write machine-aware code. Our technique can include such a processor for distributing the data analysis tasks. We chose the MASS library because it is more accessible to interdisciplinary researchers (see Section II.C for details).

B. Background on NetCDF

The internationally accepted data format of climate science data is the Network Common Data Form (NetCDF), which is a highly efficient, array based file [3]. NetCDF itself includes portable libraries for a number of programming languages and packages for scripting languages [14]. These include C, Fortran, Java, Matlab, Python, and R. While these libraries aid in the development of tools to process data in this format, among the provenance-aware tools that are available, such as NetCDF Operators (NCO) [19], provenance collection is difficult to configure and the lack of a common output format makes the available provenance difficult to use.

C. Background on MASS Library

The MASS library addresses the semantic gap between analyzing algorithms and their actual implementation [10]. The MASS library abstracts away many of the complex and technical details of parallel programming, such as inter-process communication and shared resource management [8]. For the programmer, one simply specifies the number of computing nodes and resources to use and MASS manages the intricate parallel programming details under-the-hood.

MASS relies on inter-node and inter-thread communication, but abstracts such communication from the logical design of algorithms. In other words, if a programmer has a 100 x 200 x 47 simulation space that is running across three processors and four threads per processor, the programmer does not need to know how to divide that simulation space between the processors or threads. Barrier synchronization allows parallelization to advance in lock-step removing the need to manually handle concurrency issues.

Key to MASS's abstraction is the concept of Place and Places. A Places object is an instantiated, multi-dimensional simulation space (see Figure 1). The entire picture can be thought of as a single Places object that holds a multi-dimensional grid of Place objects. Each individual square formed by the grid within the Places object is a single Place object. A Place object can have distinct values, depicted by different colors and can communicate with other Place objects, shown by the blue arrows. Place objects are logical locations within the simulation space which can host mobile agents, or execute code themselves. In our technique, each Place maintains the entire Pacific Northwest climate data in a given time segment such as 6am, 12pm, 6pm, or 0am on a given date in a given month.

MASS does not depend on a shared memory paradigm, which is used by fork-join frameworks. Instead it uses remote computing nodes, with each node maintaining an independent memory space. When a user program calls MASS.Init(), this function contacts ssh daemon processes, each running on a different remote process, establishes a secured TCP link, and asks each daemon process to launch a MASS-unique slave program called MProcess. Each remote MProcess manages different stripes of distributed arrays, (i.e., Places). When the user program calls MASS.Finish(), the function kills the remote MProcesses.

III. MOTIVATIONS AND REQUIREMENTS FOR PROVENANCE SUPPORT IN CLIMATE SCIENCE

Climate researchers spend much time and effort towards peripheral, non-climate science aspects of their analyses such as file handling and manipulation. The terabyte scale of climate model data poses immediate practical issues for the individual researcher or members of small facilities [13]. Since climate researchers generally do not have extensive computer science background, they prefer to use familiar tools to perform their analyses. Thus, they are less likely to adopt tools that require much training time. On top of this, data analysis for climate researchers is highly exploratory. They usually create scripts or code for a one time analysis.

Thus, they will not spend time to create detailed workflows since these may not be used again in the future. Finally, there is a great need for data provenance to determine the validity of analyses. Insufficient provenance gathering creates difficulties in understanding the meaning of the results, especially when there is not enough model data preserved to easily visualize the results. In order to provide provenance support for these researchers, we posit that provenance techniques and tools must satisfy the following properties: accessibility, adaptability, and scalability.

Accessibility is a property that allows as many researchers as possible the ability to capture provenance and use it for their tasks. Accessibility implies simplicity, both in the storage and retrieval of provenance. Accessibility also implies low barrier to entry, requiring minimal computing resources and minimal tool training time.

Adaptability is a property that allows researchers to capture provenance regardless of changes in their methods or processes. The ability to accommodate changes is necessary in a highly exploratory nature of data analysis in climate research. Not only do methods or processes change, but the computing environment(s) on which researchers run their analyses are also likely to change.

Scalability is a property that is often associated with the ability to accommodate large datasets or large number of computing nodes. While these are important items to consider, we also refer to the scalability of the provenance capture (in terms of performance overhead) and understandability of the captured provenance.

```

31  /**
32   * @name yourUserDefinedFunction - Sample user defined function
33   *
34   * @param args - input parameters must be wrapped in an object
35   * @return - output must be wrapped in object
36   */
37  public Object yourUserDefinedFunction(Object args) {
38      // User defined functions
39      // This is where you manipulate and share data in order to implement
40      // your data analysis algorithm.
41      return (Object) null;
42  }
43
44  /**
45   * @name initialize - Sets up the MASS library and optional provenance
46   *
47   * @param status - Real time message display for GUI
48   * @param handle - MASS places identifier
49   */
50  public static void initialize(StatusAdapter status, int handle) {
51      // Perform any necessary initialization
52      pa.provenanceMessages = new ArrayList<String>();
53      try {
54          // Provenance collection
55          status.reportMessage("Storing Metadata");
56          pa = new ProvenanceAdapter("YourAnalyticsModuleProvenance",
57              fileName[0]);
58          pa.create(provenanceEntries);
59          // Start MASS
60          status.reportMessage("Starting MASS");
61          MASS.init(MASSARGS, NPROCESSES, NTHREADS);
62          // Initialize parallel computing space (Places object)
63          pacNorthwest = new Places(handle, NAME, null,
64              fileRange);
65          status.reportMessage("Computational Nodes created");
66          // Initialize all computing nodes
67          pacNorthwest.callAll(init_, fixer);
68          status.reportMessage("Computational Nodes initialized");
69      } catch (Exception ex) {
70          pa.log(Level.SEVERE, null, ex);
71      }
72  }
73
74  }
75
76  }
77
78  }
79
80  }
81
82  }
83
84  }
85
86  }
87
88  }
89
90  }
91
92  }
93
94  }
95
96  }
97

```

Figure 2. A researcher implements a climate analysis calculation in the `yourUserDefinedFunction` method (line 37). The `initialize` method initializes the `ProvenanceAdapter`.

IV. PACIFIC NORTHWEST CLIMATE ANALYSIS TRACKER

We now discuss how PNCA Tracker achieves these required properties.

A. Achieving Accessibility through Simple Interfaces and Modular Provenance Files

We provide mechanisms for easy development of parallelizable analysis modules that are integrated with provenance collection and for the usage of modular provenance files which are stored in commonly used file formats.

The development of analysis modules is straightforward and caters to researchers who are familiar with scripts (e.g., Python) or high level programming languages (e.g., Java). A researcher simply needs to implement a few required functions (read and write methods) and user defined functions (in Figure 2, it is the `yourUserDefinedFunction` method), and to place this function in a specified location. Doing so will automatically trigger the collection of a basic set of provenance information (e.g., geographic metadata) as discussed in the next section.

In addition, we created modular provenance files to assist researchers in quickly retrieving the appropriate provenance: result-specific provenance, execution-specific provenance, and error logs. We also store these files in commonly used file formats to enable researchers to easily share them with collaborators. All the provenance files are stored in the same directory. Researchers may choose to use a file-naming convention to connect the result-specific provenance with the execution-specific provenance and error logs.

Result-specific provenance contains information regarding the data's parent files and the processing used (e.g., steps within the analytical module). It also contains the climate model used, the date of data generation, characteristics of the output data itself, date range of the arrays, and geographic metadata (i.e., information to locate the results over a map of the Earth to allow the results to be renderable by any geographic information system (GIS) viewers). Result-specific provenance is embedded within the results file within a NetCDF metadata. By co-locating the data with its provenance, we minimize the chances of the provenance being lost or inaccessible.

Execution-specific provenance, meanwhile, contains provenance related to the MASS execution state. We store execution-specific provenance in a separate NetCDF file because a single analysis can produce hundreds of NetCDF files. This scheme avoids replicating the same provenance across all the results files, and thereby minimizes the storage requirement. Embedding provenance in NetCDF files was chosen to support easy query and retrieval [9].

Finally, we store in a text file format errors encountered during an analysis execution (i.e., error logs). These errors may be associated with the MASS environment or they may be associated with the analysis modules developed by the researchers. Similar to execution-specific provenance, the error logs are stored separately. By using a non-tool-specific file format, .txt file, the errors are viewable by any member of the research team.

B. Achieving Adaptability through Provenance Customization

To provide adaptability, we use different levels of provenance customization: no customization, semi-customization, and full customization. The capabilities of each level are folded into the next level.

1) No customization

The first level, no customization, requires no human intervention. As long as the analysis module is integrated within our provenance tool support, basic provenance collection automatically occurs. This includes exceptions and geographic metadata.

If an exception is tied to an operation involving a MASS Place, the Place is also identified. In addition to the exception messages, we also include the logical coordinates where the exception occurred. For example, if the data for climate data simulation analysis file at noon on July 18th of 1974 created an exception in the analysis, we include these temporal coordinates in our error log.

Provenance is also captured through the geographic metadata. The climate data contains a large amount of metadata stored in the header file. When a file is identified to be read, the metadata is automatically extracted. When a results file is created, this information is used to create the metadata used for the header of the results file.

Of the metadata available from the climate simulation files, some are broadly applicable for every results file. This would include all of the GIS data: results cover the same geographic range as the source data. Metadata is duplicated for every results file. Other metadata cannot be copied exactly because there are multiple entities involved, such as time-based data. When the results file of a time-series analysis is created, several entities contribute to the data, each one with a distinct time slice. Appropriately transforming the metadata to reflect this is also an activity of our provenance collection at the level of no customization.

2) Semi-customized provenance collection

The second level, semi-customized provenance collection, requires some input from researchers. For example, researchers may choose to collect detailed information regarding the MASS execution state. State information can assist users in understanding the extent of parallelization employed (e.g., scalability of their analysis modules) and the source of errors. Execution state is particularly important because debugging software in a distributed system can be difficult. State information includes number of Place objects generated, number of processes, threads per process, the different processors employed, total execution time, and the analysis module used. Some of the provenance data are taken directly from the argument parameters input to the MASS engine. Others are collected from information given to the instantiation of the Places space, such as the number of logical computational nodes.

3) Fully customized provenance collection

The third level, fully-customized provenance collection, requires more specification from researchers but provides richer metadata that can encapsulate contextual information

about an analysis (e.g., hypothesis being tested). Within the user-defined analysis module, users may embed provenance information related to each analysis step or method within the module, similar to adding comments in a code where a tag or a note related to the algorithm is recorded. This provenance information is stored with the execution-specific provenance. The captured provenance is dependent on the writer of the analysis class.

Creating a fully-customized provenance collection is fairly straightforward. Users simply create a new instance of the Provenance Adapter object (see line 80, Figure 2). Once it is created, they may add provenance logs to record steps executed in the analysis.

C. Achieving Scalability through Coupling Provenance with Analysis Logic

When the operation scales to include dozens, hundreds, or even thousands of computing nodes, it is still important to collect discernible provenance. The results-specific provenance is strictly tied to the logic of the analysis. The captured execution-specific provenance is also scalable because we get an overview of the entire computation cluster and not by individual node. This is because the MASS library abstracts the physical computation from the logical model. Therefore, if we run the same analysis, going up from two nodes to two thousand nodes, we expect, other than the necessary communication overhead, a commensurate increase in efficiency without any degradation of the understandability of the provenance being collected.

V. PNCA TRACKER TOOL SUPPORT

The PNCA Tracker tool support is implemented within the PNCA Framework to facilitate parallelized analysis of climate model data. The tool is implemented in Java, to enable running on different operating systems. This section covers more details about the tool.

A. Tool Design

The overarching architecture employed is one of adapters, based on the adapter design pattern (see Figure 3). We used this pattern to loosely-couple the major modules within our system to support extensibility. Each major module of the system is treated as an “adapter” and their

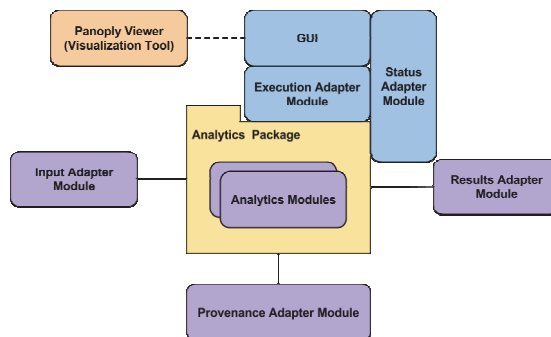


Figure 3. Design diagram of the PNCA Tracker tool support. Modules connected with lines are loosely coupled.

main purpose is to provide “dependency sinks.” Changes in how each adapter operates should only result in internal changes to the adapter without affecting their interfaces to the rest of the framework.

We developed an integrated system with a graphical user interface (GUI) and several modules that can be applied to read data, record results, and supply important provenance information. These input and output tasks are able to be performed in parallel on multiple threads and processes through the MASS library, which resides within the Execution Adapter Module.

The Analytics Package contains the analysis modules developed by researchers and function as “plug-ins” to the tool framework. Users may write an unlimited number of functions to carry out an analysis. The user defined functions are performed in parallel and may contain customized instructions on provenance collection (see Figure 2). The researcher implements methods to read, write, and analyze climate simulation data (e.g., `yourUserDefinedFunction()` shows the minimal format of such a method). The `initialize()` method (line 74 in Figure 2) is called by the Execution Adapter module to allow the analysis to be parallelized by MASS.

The GUI, Status Adapter, and Execution Adapter modules are used to execute the analysis packages. The GUI allows for the selection of climate data files and specific analysis modules. The analysis modules are automatically added to the dropdown when the users save the files in the proper location. The Configure Input command allows selection of source files. The GUI also allows researchers to view the results of the analysis directly in a viewer. We currently integrated our tool with the Panoply Viewer. The Status adapter transmits execution-time status messages from the analysis module to the GUI. Most important is the Execution Adapter, which controls the parallel execution of the analysis module across multiple processes and threads using the MASS engine. The Execution Adapter is also responsible for capturing execution-specific provenance.

```
File "FrostTally1969"
Dataset type: NetCDF-3/CDM
netcdf file:/home/brett/NetBeans%20Projects/pac-nw-climate-analysis/FrostTally1969 {
// The dimensions in the output file. This geographic area representing the Pacific Northwest is
partitioned into a 162x123 grid.
dimensions:
  Time = 1;
  num_metgrid_levels = 11;
  DateStrLen = 19;
  west_east = 162;
  south_north = 123;
  soil_layers_stag = 4;

// The variable(s) holding the results of the analysis, in this example a count of frost days for the
year 1969.
variables:
  int FrostDays(south_north=123, west_east=162);

// Global attributes contain source model and geocoding data provenance.
// global attributes:
:TITLE = " OUTPUT FROM WRF V3.1.1 MODEL - ON PRES LEVELS";
:START_DATE = "1968-12-30 00:00:00";
:SIMULATION_START_DATE = "1968-09-01 00:00:00";
:WEST-EAST_GRID_DIMENSION = 163; // int
:SOUTH-NORTH_GRID_DIMENSION = 124; // int
:BOTTOM-TOP_GRID_DIMENSION = 11; // int
```

Figure. 4. A partial list of metadata copied from the original source files and applied to the resulting output files.

B. Tool Implementation using the MASS Library

We use the MASS library for the backbone of execution of the custom analysis modules developed by climate researchers. We use two major methods provided by the library: `ExchangeAll` and `CallAll`. `ExchangeAll` reveals information, providing for communication between Place locations within a Places space, and between multiple Places spaces. `CallAll` performs the work, running code, allowing up to all Place spaces to perform calculations. Places have an accessible, machine-unaware indexing structure that allow specific subsets of Place objects to be identified, meaning that groups of Place locations can execute code while others remain quiescent.

When researchers create their own analysis modules, these modules extend the Place object, so its methods can be run across large quantities of data in parallel. By using specific Place objects identified from their index values, data can be quickly reduced. For example, in an analysis module that calculates the number of days when the temperature dropped below freezing for each geographic cell of climate output, the multiple input files would be reduced to a single results file of the time-series analysis.

C. Usage Scenarios

We built our tool support with the following usage scenarios. Researchers capture provenance from an exploratory analysis so that they can retrace their steps if their analysis yields useful results. Analysis developers create a reusable set of analysis tools that research students can later run and record provenance. After analysis execution, analysis developers fine-tune their algorithms based on the MASS state information. Prior to publication of results, developers consult the results file(s) (which contain result-specific provenance), the execution-specific provenance, and error log to check their results.

VI. EVALUATION

We evaluated PNCA Tracker using provenance queries it can answer, a case study on local climate data analysis, and timing results. We also discuss limitations of our technique.

A. Provenance Queries

1) *What input climate data files were analyzed to produce FileXYZ.nc?*

This can be answered by referring to the global attributes within FileXYZ.nc. As shown in Figure 4, the three global attributes: `TITLE`, `START_DATE`, and `SIMULATION_START_DATE` identify the specific parent file(s) and time slice. Figure 4 also shows that the resulting output files include the necessary geographical information to allow results files to be visualized with GIS viewers. This provenance information is collected at the no customization level.

2) *What analysis steps were taken to produce FileXYZ.nc?*

Detailed processing steps within the analysis module, if specified by the analysis developer, are captured and saved with the execution-specific provenance.

B. Local Climate Research Case Study

We asked a climate impacts researcher to evaluate PNCA Tracker’s usability. Using the tool, s/he was able to evaluate data from the Pacific Northwest derived from the Weather Research and Forecasting (WRF) Model.

WRF is a state-of-the-art mesoscale numerical weather prediction system designed to serve both operational forecasting and atmospheric research needs [4]. This model has been developed and used extensively in recent years for regional climate simulation [17]. WRF has been implemented as a regional climate model over the Northwest United States at 12 km grid spacing [16].

The dataset covered climate models from 1980 spanning 6 hour intervals. Each time slice was around 56 megabytes in size. The geographic space simulated consists of a 123 by 162 grid, where each cell is approximately 15 kilometers on a side. The space is extended into the third dimension through a number of atmospheric layers above the ground, and soil layers below the ground. Each cell contains 150 variables of measurement, such as temperature and pressure.

The study was run with semi-customized provenance collection, which provides metadata from the climate data examined as well as clustering information from the analysis. The provenance information enabled the researcher to verify the climate models’ data.

Concerning accessibility, it took the user less than 30 minutes to understand the tool and to analyze some sample data with a predetermined calculation for Simple Precipitation Index. S/he also stated that the time spent recording provenance and analyzing the provenance was highly acceptable. S/he also commented that the tool “is fast. The parallel computing is great.” Since the user was unfamiliar with Java, s/he was unable to use the full provenance customization feature of the tool.

C. Timing Results

With regards to scalability, we compared the performance overhead imposed by our technique with NCO (see Figure 5). The timing analysis was run on a Linux Mint 15 (Olivia) x64 with Intel Core CPU T5600 @ 1.83GHz and 2 gigabytes (GB) of random access memory (RAM). Our tests involving minimal parallelization (two threads on a single processor) nevertheless produced substantial improvements when compared to an unparallelized version of the same algorithm and running the algorithm using NCO

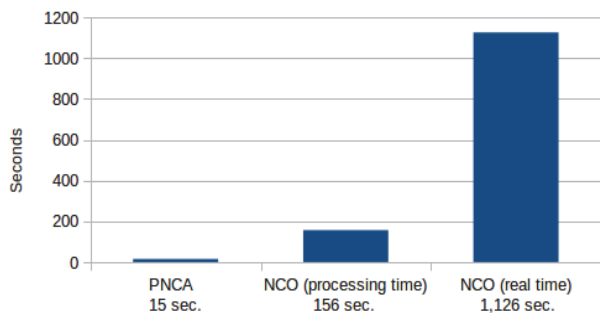


Figure 5. PNCA Tracker shows minimal performance overhead with provenance collection compared to the provenance-aware mode of NCO.

in provenance-aware mode. A moisture flux calculation using PNCA Tracker on the dataset used in the case study took 15 seconds of real time to accomplish, including the provenance capture, while running the calculation using only one variable pair (PNCA Tracker used two) with NCO took 156 seconds of processing time, and over 15 minutes of real time to complete. The PNCA Tracker was executed at the level of “full customization,” the most comprehensive provenance generation in our tool. The improvement can be attributed to the fact that PNCA Tracker first loads all the data to memory so that the calculations are highly efficient. Furthermore, the scalability that PNCA Tracker achieved due to the MASS engine’s ability to distribute as well as parallelize the analysis is far superior to NCO.

We also compared the overhead imposed by customizing provenance (see Figure 6). The dataset used for the timing evaluation was an entire year of WRF climate models from 1970 (10.27 GB total file size). The timing analysis was run on a Red Hat Enterprise 5.11 (Tikanga) with Intel Xeon E5520 @ 2.27 GHz and 6GB of RAM. Three sets of 5 runs each (total of 15 runs) were performed, with the different order of runs for each set. The average runtime of a no customization run was 8.922 seconds, semi-customization was 8.910 seconds, and full customization was 9.241 seconds. This indicates that fully customizing provenance imposes minimal additional overhead compared to no customization.

D. Limitations

There are substantial overhead requirements involved with the use of the PNCA Tracker. Each Place object is a complex data structure that takes up large amounts of memory. With a commodity laptop and 2GB of RAM, the maximum size for an analysis is about three months’ worth of six-hour time slices. However, the point of the inclusion of MASS is to allow these calculations to be performed over a computer cluster, allowing for immediate scalability.

The MASS library abstracts away the parallel and distributed nature of the program execution from the developer. One significant drawback is that we are not collecting provenance on the location of execution (i.e., processing node) for a subset of Place objects. It is possible to collect this provenance information and we are currently examining how this can be efficiently achieved with minimal overhead.

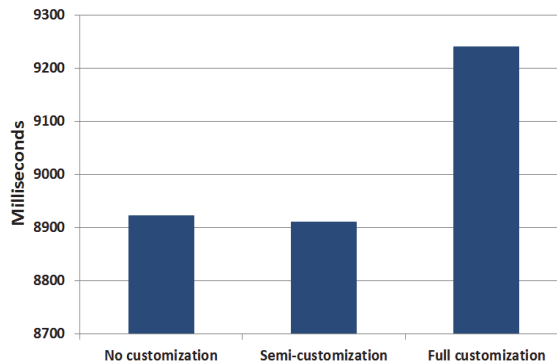


Figure 6. PNCA Tracker shows a performance overhead of less than half a second for fully customizing provenance.

VII. MAPPING PROVENANCE CONCEPTS TO W3C PROV

W3C PROV is a community standard for representing provenance [2]. The three main concepts in W3C PROV, entities, activities, and agent, have direct mappings to concepts in PNCA Tracker, as shown in Table 1.

The entities in PNCA Tracker are the input climate data and the results files (output), both in NetCDF format. Although our area of focus is the Pacific Northwest region of the United States of America, southern Canada, and the northeastern Pacific Ocean, the PNCA Tracker can read and apply its analysis package to any properly-formatted climate data. This is due to the universal agreed-upon format for climate data in terms of variable names and data types.

The activities can be mapped to the steps within the analysis algorithms and the interactions with the tool's user interface (e.g., selection of analysis module, selection of input files).

The agent can be mapped to multiple concepts in PNCA Tracker. Researchers who create the analysis modules can be considered an agent. The MASS Environment, Execution Adapter Module, external climate data simulation programs, and result viewers can also be considered agents.

Analysis modules are unique in that they can be mapped to these three concepts depending on time and perspective. During and after an analysis module is written, it is considered an entity that an activity (writing) produced which is associated with an agent (a researcher). During a module's execution, it itself can be considered an agent by which the output files are associated. After a module's execution, the individual steps within the analysis module can be considered activities that generated output analysis files.

VIII. CONCLUSION AND FUTURE WORK

The PNCA Tracker is an improvement over the existing state of the art for climate researchers who wish to collect provenance but have limited access to, or training with, HPC or specific platforms or technologies. PNCA Tracker is accessible to researchers with minimal computer science background, adaptable to different types of climate analysis and different operating systems, and scalable to large computing clusters. Our technique stores captured provenance in a modular fashion, to minimize access time and to minimize the storage requirements. Our technique also allows researchers to customize the provenance collection. Our set of evaluations (provenance queries, climate research case study, and timing results) indicates that PNCA Tracker is a feasible option to provenance tracking in a parallel and distributed environment. Finally, provenance concepts in PNCA Tracker can also be mapped to concepts in the W3C PROV.

A future goal is to use agent analysis. As the name implies, MASS supports the deployment of large quantities of mobile agents into its environment. Two immediate uses for such agents would be an improvement to the basic time detection algorithm and meteorological phenomena detection, such as locating atmospheric rivers or storms. We also plan to export provenance into the W3C format.

TABLE I. PROVENANCE MAPPING BETWEEN W3C PROV AND PNCA TRACKER.

W3C PROV	PNCA Tracker
Entities	Climate data in NetCDF format (input) Results file(s) in NetCDF format (output) Analysis modules
Activities	Analysis module steps Interactions with the PNCA Tracker GUI
Agent	Analysis module developers Climate data simulation programs Execution Adapter Module MASS Environment Viewers (e.g., Panoply) Analysis module (execution)

ACKNOWLEDGMENT

The authors would like to thank the following for their active assistance and participation: Dr. William Erdly for academic project guidance, Delmar Davis for provenance collection advice, the MASS Project Team for advice and guidance concerning the MASS Library, and the student cohort of the UWB Master's Project from Summer and Autumn 2013 for peer reviews and advice. This work is based in part upon work supported by the US National Science Foundation under Grant No. ACI 1350724. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

REFERENCES

- [1] Message passing interface (MPI) standard. <http://www.mcs.anl.gov/research/projects/mpl>. retrieved: Jan. 2015.
- [2] PROV-Overview. <http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/>. retrieved: Jan. 2015.
- [3] Status of standards body endorsements of NetCDF and related conventions. <http://www.unidata.ucar.edu/software/netcdf/docs/standards.html>. retrieved: Jan. 2015.
- [4] The weather research & forecasting mode. <http://www.wrf-model.org/index.php>. retrieved: Jan. 2015.
- [5] Message Passing Interface Forum. MPI-2, chapter Extension to the Message-Passing Interface, Chapter 9, I/O. University of Tennessee, 1997.
- [6] S. Akoush, R. Sohan, and A. Hopper, "HadoopProv: towards provenance as a first class citizen in MapReduce," Workshop on the Theory and Practice of Provenance, Apr. 2013, pp. 11:1-11:4.
- [7] Y.-W. Cheah, R. Canon, B. Plale, and L. Ramakrishnan, "Milieu: Lightweight and configurable big data provenance for science," Proc of Int'l Congress on Big Data, Jun. - Jul. 2013, pp. 46-53.
- [8] T. Chuang and M. Fukuda, "A parallel multi-agent spatial simulation environment for cluster systems," Proc of Int'l Conf on Comp Science and Eng, Dec. 2013, pp. 143-150.
- [9] D. B. Davis, H. U. Asuncion, G. M. Abdulla, and C. W. Carr, "Towards recovering provenance with Experiment Explorer," Proc of Fifth Int'l Conf on Info, Process, and Knowledge Management, Feb.-Mar. 2013, pp. 104-110.
- [10] J. Emau, T. Chuang, and M. Fukuda, "A multi-process library for multi-agent and spatial simulation," Proc of Pacific Rim Conf on Communications, Computers and Signal Processing, Aug. 2011, pp. 369-375.

- [11] E. Santos et al., "Designing a provenance-based climate data analysis application," Int'l Provenance and Annotation Workshop, Jun. 2012, pp. 214–219.
- [12] J. B. Buck et al., "SciHadoop: Array-based query processing in Hadoop," Proc of Int'l Conf for High Performance Computing, Networking, Storage and Analysis, Nov. 2011, pp. 66:1–66:11.
- [13] J. Gray et al., "Scientific data management in the coming decade," SIGMOD Record, vol. 34, no. 4, Dec. 2005, pp. 34–41.
- [14] J. Li et al., "Parallel NetCDF: a high-performance scientific I/O interface," Proc. of Supercomputing, Nov. 2003, page 39.
- [15] D. Goodman, "Provenance in dynamically adjusted and partitioned workflows," Proc of the Int'l Conf on eScience, Dec. 2008, pp. 39–46.
- [16] E. P. Salathe Jr, L. R. Leung, Y. Qian, and Y. Zhang, "Regional climate model projections for the State of Washington," Climatic Change, vol. 102, no. 1-2, May 2010, pp. 51–75.
- [17] L. R. Leung, Y.-H. Kuo, and J. Tribbia, "Research needs and directions of regional climate modeling using WRF and CCSM," Bulletin of the American Meteorological Society, vol. 87, no. 12, Dec. 2006, pp. 1747–1751.
- [18] R. Stevens, J. Zhao, and C. Goble, "Using provenance to manage knowledge of in silico experiments," Briefings in Bioinformatics, vol. 8, no. 3, May 2007, pp. 183–194.
- [19] C. Zender, H. Butowsky, and W. Wang, Welcome to the netCDF Operators (NCO) homepage. <http://nco.sourceforge.net/>. retrieved: Jan. 2015.
- [20] D. Zhao, C. Shou, T. Malik, and I. Raicu, "Distributed data provenance for large-scale data-intensive computing," Proc of Int'l Conf on Cluster Comp, Sep. 2013, pp. 1–8.

Discount Coupons Dematerialization: a Comprehensive Literature Review

Gonçalo Paiva Dias
ESTGA / GOVCOPP
University of Aveiro
Águeda, Portugal
gpd@ua.pt

Hélder Gomes
ESTGA / IEETA
University of Aveiro
Águeda, Portugal
helder.gomes@ua.pt

Jorge Gonçalves
ESTGA / GEOBIOTEC
University of Aveiro
Águeda, Portugal
luisjorge@ua.pt

Daniel Magueta
ESTGA
University of Aveiro
Águeda, Portugal
dmagueta@ua.pt

Fábio Marques
ESTGA / IEETA
University of Aveiro
Águeda, Portugal
fabio@ua.pt

Ciro Martins
ESTGA / IEETA
University of Aveiro
Águeda, Portugal
ciro.martins@ua.pt

Mário Rodrigues
ESTGA / IEETA
University of Aveiro
Águeda, Portugal
mjfr@ua.pt

Jorge Araújo
Saphety
Lisboa, Portugal
jorge.araujo@saphety.com

Abstract— This article presents a comprehensive literature review regarding digital coupon processing in its various aspects: suppliers, retailers and customers. Current standards, solutions and platforms available in the market and proposed by the scientific community (research, patents, etc.) are presented. A brief summary of the major trends in digital coupon processing is also presented. By resuming the state of the art in digital coupon processing, the article may be useful both for researchers and practitioners interested in the topic.

Keywords— *Digital Coupons; Coupons Management; Redemption; Automatic Clearing; GSI; e-Commerce*

I. INTRODUCTION

Typically, a coupon is a certificate that allows a consumer some sort of incentive to buy a product or service. Although the incentive is usually a reduction in price, coupons can also be used for reimbursements, combined offers, free samples, or other types of promotions (e.g., sweepstakes or contests) [1]. The economic and financial crisis has reinforced the concerns of consumers in relation to issues of savings, causing a change in their behavior, particularly in relation to the savings obtained via coupons. The market share of digital coupons is increasing. Given the pace of change in mobile communications, they present enormous opportunities for companies with strategic vision that can use them to attract and retain customers [2]. Online

social coupons that offer daily deals are gaining relevance. On the one hand, due to the emergence of new distribution channels, such as email and mobile applications, but also because they are reaching new markets and products (e.g., luxury goods, medical and dental care, etc.). Revenue from discount sites like Groupon and Living Social reached the 2.67 billion US dollars in 2011, representing an increase of 138% over the previous year [3]. According to NCH, manufacturers of packaged products distributed over 311 billion dollars in coupons in 2009, and enabled consumers to obtain total savings of 3.5 billion [4].

The development of information technology and telecommunications generated a growing interest in mobile marketing. In the late nineties, market analysts projected a great future for marketing initiatives that used mobile devices [5]. Mobile coupons can be sensitive to time and/or location. Firms may use time sensitive coupons when sales are low due to the time of day or due to business seasonality. Often, companies have a deep knowledge of their customers, allowing them to create profiles of their tastes and needs, and use coupons to meet their specific needs [6].

For consumers, social coupons proved to be an excellent tool to support purchases. Besides offering great discounts, they allow consumers to have a first experience with various products at a lower price. Offering services that allow the use of online and/or mobile coupons has proved to be a

profitable business model. However, the same is not necessarily true for those that offer the coupons themselves. In effect, big discounts, higher redemption rates when compared to traditional coupons, and payments to service providers makes it extremely difficult to achieve profitability in the short term. Even in long-term profitability is uncertain [7]. Given these concerns, companies hesitate to develop large-scale initiatives and are cautious about the potential of the coupons. Despite these fears, some prospect a promising future for these strategies, since both coupons technology and the propensity of consumers to use it are evolving [8].

The remaining of this article is organized as follows: in Section 2, existent standards are introduced; in Section 3, current platforms and solutions are resumed; in Section 4, patents are presented; and in Section 5, future trends are addressed.

II. STANDARDS

One of the reasons why paper based coupons persist to be the predominant solution for processing promotional offers is the fact that there are well-defined and accepted methodologies for their generation. Their processing can easily be integrated in both suppliers and retailers systems. When digital coupons are at stake, in particular when its distribution is made through mobile devices, no similar and standardized methods exist.

Interoperability issues between different technologies for processing digital coupons contribute to the lower redemption rates of this type of coupons when compared to the traditional ones in paper format. Most studies indicate that digital coupons redemptions account for only 2% of total redemption [9]. Points-Of-Sales (POS) are the main bottleneck, because of the heterogeneity of systems and versions and because most of them are unprepared to read bar codes from mobile devices. To overcome these obstacles, it is necessary to have standards that enable compatibility of various systems in the different stages digital coupons processing: from generation and communication to redemption and ransom and financial reconciliation.

The Mobile Marketing Association (MMA), the premier global non-profit association representing the various actors of the mobile marketing arena, seeks to define and establish standard languages and platforms for processing various types of promotional offers such as coupons, incentives, rewards, etc. It created the Mobile Coupon Ad Unit (MOCAUS) committee with this purpose. As the result of this effort, in [9] it is presented and described a processing platform that standardizes the various stages of the process (Figure 1). In [10], it is provided a set of diverse documents such as studies, reports, and guides to best practice in mobile marketing.

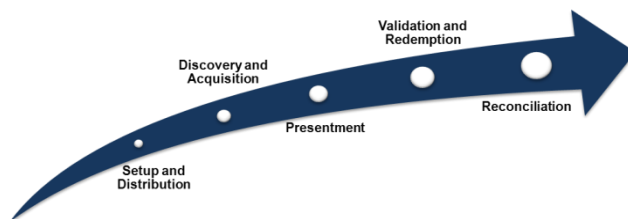


Figure 1. Platform for processing digital coupons proposed by MMA (adapted from [9])

Relating the standardization of business processes associated with the use and treatment of coupons, GS1 [11] is the main worldwide organization seeking to define and establish open standards for coding and managing the flow of goods, services and data through the value chain [12]. The set of open standards of the GS1 are recognized by the International Organization for Standardization (ISO) and allow the correct identification (national and international) of items (products and services), logistic units, and commercial actors across the value chain and activity sectors. The identifiers of the European Article Numbering-Uniform Code Council (EAN•UCC) can be represented by barcodes and consist of three elements: Global Trade Item Number (GTIN) -; Serial Shipping Container Code (SSCC) -; and Global Location Number (GLN). Apart from the unique identification, these codes enable the exchange of additional information such as expiration dates, serial numbers, lot numbers, etc. [13].

More specifically, with regard to the processing of digital coupons itself, the normative document [14] from GS1 establishes a set of specifications that define the first version of the ‘Digital Coupon Management standard.’ In [15], it is presented a first proposal for standardizing the digital coupons management process (Figure 2). This proposal specifies standards for the management process, the identification of objects and the data model.

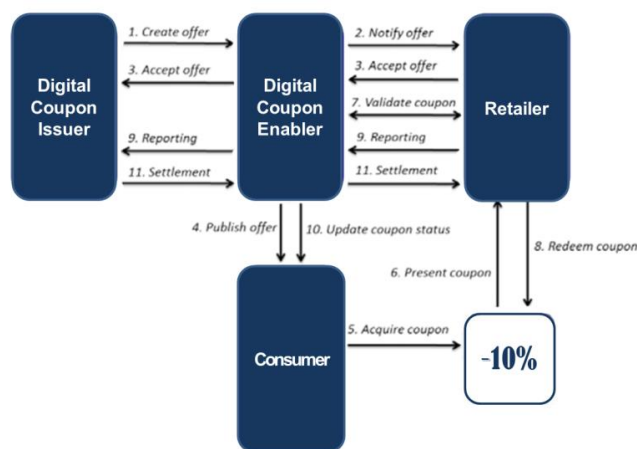


Figure 2. Process management of digital coupons as proposed by GS1 (adapted from [15])

Relating the standardization of data communication, GS1 Global Data Synchronization Network (GDSN) [16] defines a set of standards that specify the connection of the different

actors in the value chain with the GS1 Global Registry through a network of certified data repositories (GDSN-certified data pools) (Figure 3). In this network, all items are identified by a unique combination of GTIN and GLN identification codes.

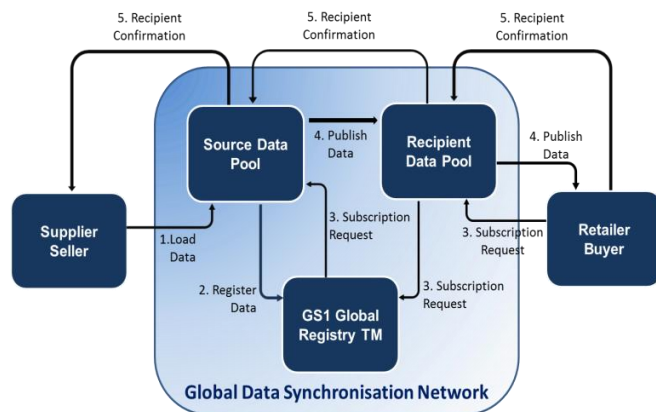


Figure 3. GS1 Global Data Synchronisation Network (adapted from [16])

Inefficiency and fraud rate related to coupon processing in paper format results mostly from the fact that many retailers fail to make a correct validation of coupons offers at the POS. Therefore, the most logical way to solve the problem is to block the redemption of invalid coupons when presented at the POS by consumers. This can be achieved through the combination of three factors [17]: the new GS1 Databar barcode; retailers' systems compatible with GS1 standards; and ultimately, a real-time validation service interconnecting POS systems and the data pool server(s) containing date information on coupons offers.

Despite GS1 being better known because of the barcodes used by companies to identify their products, their standards encompass other aspects than coding. In effect, the GS1 system of standards enables organizations to identify, extract and share information in the supply chain [18]. More information on the architecture of the GS1 standards system can be found in [19].

III. CURRENT PLATFORMS AND SOLUTIONS

The new digital communication channels are an excellent opportunity for the various stakeholders involved in the cycle of buying and selling products: suppliers, retailers and consumers. In particular, the use of coupons in digital form as well as its processing by electronic means offers the opportunity for suppliers and retailers to eliminate most of the problems associated with the traditional coupon processing cycle in paper. There are several companies that provide solutions for the different stages of digital coupons processing: issuance, distribution, validation, reconciliation and payment. However, these solutions often involve proprietary and local implementations based on non-standard protocols. They can be classified in two major types: global solutions; and partial solutions.

A. Global Solutions

In general, this type of platform relies in a base architecture that allows processing solutions for both digital and non-digital coupons, with a very similar flow of information. From the consumer's viewpoint, the available systems can be used to combine discounts from the supplier with the local retailers. To use the service consumers register on the platform and add their coupons to their account. The coupons can be added from a mobile device or from a site (site of the platform itself, retailers sites, coupon distributors sites, etc.), with a pre-defined limit. Customers must have a mobile number or one or more loyalty cards associated with your account. The mobile device number or any of the cards associated with the account can then be used to redeem the coupons for the purchase of products at the POS. For some platforms consumer may print the coupons on paper to present them in the POS. The management of loyalty cards and mobile phone numbers is performed by the customer himself. He may add, change or remove any of the items associated with its account. Coupons are automatically removed from the consumer list when they are used at the POS. Besides automatic removing, some platforms allow consumers to remove coupons themselves from their account. This option allows consumers to replace a coupon when they find a more advantageous discount voucher for a specific product, for example.

From the suppliers and retailers viewpoint, they offer coupons on the platform, which automatically updates the information in their website(s), twitter profiles, social networks and main search engines, thus informing consumers (through custom promotional marketing campaigns and taking into account the profile of the specific consumers). Concerning customized marketing, platforms provide some applications for mobile devices, which using geotagging technology enable consumers to receive alerts on the existence of offers of coupons when they are physically close to their favorite retail stores. In this case, both the activation of the alert service and its configuration (conditions and preferences) are made by consumers themselves. Whether distributed by the Web, email, newsletters, videos, social sharing, mobile devices, or in the retailer's stores (for example, by capturing 2D codes), all digital coupons offers are processed electronically at the time of submission by the consumer at the POS.

B. Partial Solutions

In addition to the global solutions that deal with the entire cycle of coupons processing, many other solutions seek to meet the needs of retailers in terms of promotional campaigns, including discount vouchers (coupons) associated to specific products and other types of promotions that entitle discounts on services, restaurants, bars, clothing and convenience stores, etc. Promotions are offered in a similar way to the global solutions platforms. For consumers who have mobile devices there are applications that allow downloading of digital promotions. For other consumers, promotions are sent via services like Short Message Service (SMS), Multimedia Messaging Service (MMS), etc. To benefit, consumers only have to show the coupon at the place

of purchase of the product or service. Promotions are made available to consumers according to their consumption habits. Nevertheless, they always have control over the information they want to receive. Some of the systems available on the market also provide geolocation features that allow showing the consumer where he can benefit from promotions and which places near its location provide those promotions.

Relating coupons processing in Portugal, PacSis (Systems Promotion and Marketing) [20] provides the service of managing discount vouchers. One of its newest marketing solutions is offering online coupons in partnership with Coupons.com Incorporated [21], also including mobile marketing and profiling targeting for brands. However, the solution does not currently include the integrated electronic processing of the different phases of digital coupons deployment. Digital coupons must be printed for later redemption at retailers.

The ability of real-time analysis of data on purchases and redemption of coupons is one of the increasingly important features for those responsible for the definition of promotional campaigns. Several companies seek to explore and develop platforms to provide this kind of analysis based on the buying cycle, allowing the recommendation of products and the customization of coupon offers to consumers.

In [22], it is presented a platform for implementing discount coupons which seeks to maximize both customer satisfaction and business profitability. The platform is based on three components: demand and recommendation of coupons; distribution of digital coupons; and multimodal redeem of coupons (1D barcode, 2D matrix codes, alphanumeric codes, and other visually represented codes).

C. Research Work

With respect to research-oriented view on the topic, the incorporation of Near Field Communication (NFC) technology in processing systems of digital coupons is currently a hot research topic. In [23] and [24], two different works present the WingBonus system, a solution used for dissemination, distribution, validation and management of vouchers, loyalty cards, and all kinds of coupons using NFC technology. The issues of security and usability are discussed in [25], where the authors suggest a vouchers management system that integrates NFC technology. A practical offline payment system based on digital vouchers using NFC in mobile phones is presented in [26], a project that assesses the feasibility of such a system, from a technical and security perspective, using tangible NFC devices.

Another important field of research is related to security concerns where some interesting works are being developed ([26] is a good example). In [27], the authors propose a chaotic maps-based authentication scheme for e-coupon systems that satisfies security and functionality requirements while preserving efficiency. A system of virtual coupons that is protected against illicit use is made in [28], where the authors provide a shortlist of possible attacks and describe the protocol to prevent them and the requirements for all major components. In [29], the authors propose a new

efficient and secure micro-payment scheme, named e-coupons, which can provide the users the facility for delegating their spending capability to other users or their own devices.

Another important line of research is related to the utilization of mobile devices. In [30], a study is made on omnichannel commerce and on how mobile affects in-store traffic and sales, and in [31], a report makes an analysis on how coupons offers are reacting and adapting to mobile. A solution, called Mobeam, to overcome the problem of the inability of most barcode readers at retailers to reliably read a barcode displayed on phones, and through that to promote further utilization of mobile phone as a mean of coupon utilization [32].

With the increasing development of geolocation technologies, this field is becoming a central topic in digital coupons research. In [33], the authors made a study on location-based advertising on mobile devices and social networking that use local-tracking technology to target clients. A study on the e-coupons strategy problems in Location Based Advertising (LBA) using a full information model in a reduced optimization problem is made in [34].

IV. PATENTS

Various forms and solutions for processing digital coupons are available in the market. Many of them are proprietary solutions that do not comply to standards. Amongst them, several successful applications explore new technologies and tools available in the area of wireless communications. Moreover, the new standard for digital coupon - GS1 DataBar - fostered the development of solutions capable of overcoming issues related to interoperability between systems [17].

Over the past four years, with the advances in terms of mobile communications, many companies proposed and patented systems for the electronic treatment of different stages of digital coupons processing. In [35], Coupons.com Incorporated describes a set of techniques and mechanisms for generation, distribution, redemption, reconciliation and payment of coupons. In the proposed architecture, a distribution of coupons entity, allows that, through a server, a set of previously registered entities can generate promotional coupons offers. Through a network, the distributor server receives from retailers the information regarding retail coupons presented for redemption by consumers. In response, the server of the distributor determines their validity, checking not only the terms of the offer, but also if they have been previously redeemed in other retailer, thus eliminating possible frauds and errors. If coupons are valid, the server labels them as redeemed causing the retailer to be credited for the amount of the respective discount. This solution includes the possibility of having a server at the retailer, enabling that the generation of coupons for a particular offer available in the distributor's server be made during the checkout process at the retailer. The solution also enables consumers to add coupons to their account in the distributor's server or to print them in paper. Upon checkout, consumers can thus redeem coupons in various forms: in

paper; on a mobile device; or by indicating the identifier of their account in the distributor platform

In [36], a solution similar to the above is presented. However, it does not include the possibility of generation coupons at the retailer side nor printing coupon in paper. The solution includes a component for managing the distribution, redemption, reconciliation and payment for retailers, and billing and payment to suppliers.

In systems in which the consumer must provide his mobile device so coupon codes can be read, several problems arise. Namely, the processing time required at the POS checkout. The method described in [37] allows that the distribution, redemption and reconciliation be made by transmitting data of the digital coupon directly to consumer's mobile devices via wireless communication. Similarly, it is detected directly from the consumer device the existence of coupons selected for redemption, the data being also received via wireless communication.

Another area of development is the distribution of coupons and other offerings through consumer devices in direct connection with products and services. In [38], it is presented a system and respective method for the selective distribution of digital coupons based on the consumer geographical position relative the location of the retailer's store.

Another example of solution for distributing coupons is the use of Radio Frequency Identification (RFID) tags at various locations inside the shops as a means of sharing information about products, and download of coupons and other incentives, serving also as a way to detect the presence of consumers in the shops and facilitate commercial transactions in the POS. Labels placed on placards and ads near the products, shops entrances or POS can be read by consumer's mobile devices using software for detection, reading and subsequent decoding of the information provided. Nokia Corporation [39] proposes a solution for processing promotional information using RFID tags. The proposed platform enables the capture of promotional information provided by the retailer, allowing the consumer to select coupons via the mobile device for later redemption. The platform then makes the validation of redeemed coupons, showing in the device information related to its validation that is used for confirmation by the retailer. A similar solution is described in [40] by Coupons.com Incorporated. The consumer has the opportunity to get promotional information not only through RFID tags but also using Quick Response (QR) codes.

Still concerning the distribution of coupons, Apple Incorporated presents in [41] a specific solution for storage, management and redemption of digital coupons through a mobile device. Storage can be done either in a server accessible from the mobile device as locally on the device itself. Consumers enter the identifying code of the coupon or read a QR code to store coupons in the mobile device. The solution allows the device to alert the consumer of redemption possibilities according to the coupons it stores. That is, when the consumer is near a store where coupons can be redeemed, by using geotagging mechanisms or when the expiration date approaches. Additionally, alerts can be

generated whenever the device verifies that there are nearby products of the consumer purchase list for which it holds discount coupons. Checkout redemption can be made either by introducing the coupon code in the POS or automatically, if the mobile device can be used for payments using NFC.

Finally, regarding the generation of coupons and their identifications, most of the solutions that have been proposed implement coding schemes that lead to the standards proposed by GS1, particularly in regard to the designated 'mobile coupons', that follow the GS1 DataBar standard [42]. In [43], it is described a system for generating coupon offers and respective barcodes. It enables the user (supplier or retailer) to generate coupons by selecting the type of barcode. In response to this choice, he is presented with a specific interface through which he will provide the data needed to generate the coupon. In the range of possible types of coupons, the progressive discount coupons, whose value increases with the number of consumers that perform a certain action, or with approach of the expiration date, constitute an increasingly important option in terms of marketing campaigns, especially in social networks [44]. In [45], it is described a specific methodology for the generation of variable-value coupons. In this case, besides the expiration date, the coupon includes its own schema associated with data associated to its progressive value. Thus, this variable schema depends on the time interval between the generation date and the date of redemption of the coupon.

V. CONCLUSION

According to the statistics for the third quarter of 2012 presented by NCH, the issuance of food coupons decreased 3.5% when compared to 2011. However, during the same period, the coupons issued in products and services in healthcare and beauty increased significantly (10.4%). Regarding the type of emissions, the coupons are still mostly distributed in newspapers (89.7%). Digital formats represent only 2%. Lower than coupons distributed at stores (3.9%) and in magazines, mail or packaging (4.6%).

In the same report, the NCH states that "marketers have further suppressed the attractiveness of their coupon offers with less savings and less time." In fact, between 2011 and 2012 there was a decrease in the average discount and the expiration period, and an increase in coupons which required the purchase of more than one product. The attractiveness decrease "has significantly reduced the total number of coupons redeemed so far in 2012, halting a three-year growth trend." Between 2008 and 2011 the redemption of coupons grew 34.7%, but in 2012 the decrease was significant (17%). However, it is expected that in the future companies (re)start to be more aggressive in the use of coupons and in this context digital has a great room for growth, due to its current marginal share [46].

In a specific study on mobile coupons in 2013 [47], several experts predict that by 2016 consumers will redeem approximately \$ 43,000 million US dollars worldwide, compared to 5,400 million in 2011. While most companies are still reluctant to adopt this strategy, mobile coupons will increasingly be an excellent way to implement advertising

and customer loyalty campaigns without incurring in extensive advertising or other brand building methods costs.

In the same study, the main trends for the industry of mobile coupons are addressed:

- Location-based coupons - Walgreens can send coupons that apply to the store where the customer is, attracting him to a specific location.
- Barcode replacement by NFC - NFC is an intelligent technology far superior to existing methods and their application will be increasingly diverse. For example, integration with the "check-out" will allow customers to quickly pay, redeem coupons, etc.
- Dissemination of interest via social networks - It is rare to find someone who does not have a profile on a social network. By 'forcing' the publication of coupons in profiles, companies will be able to increase the interest in certain stores and products.
- Mobile Coupons as catalysts for online shopping - Customer interest in a specific product can be created by inducing him to search for it via online coupons. The placement of a link in the coupon can make the experience of purchasing online more friendly to the buyer, removing the inconvenience of looking for the article in the online store and increasing the propensity to buy immediately before the offer expires.

In its annual report, Inmar [48] considers that the decline in coupon redemption in 2012 is explained by the divergence between consumer preferences and offerings. However, consumer interest in coupons is extraordinarily high, especially among young men. These groups never sought so much for discounts, and the new tools they have at their disposal opens excellent prospects for the use of coupons.

Finally, Small Biz Trends [49] indicates that the main trends for the industry of the coupons will include the growth of localized offers, increase of online trading, more resources and tools for business users, and greater flexibility for consumers to personalize their offers.

ACKNOWLEDGMENT

This work is part of the Value4Coupons project, a QREN project (023160), co-funded by COMPETE and FEDER.

REFERENCES

- [1] D. Schultz, W. Robinson, and L. Petrison, "Sales Promotion Essentials: The 10 Basic Sales Promotion Technique and How to Use Them", Chicago, NTC Business Books, 1998.
- [2] K. Jung and B. Lee, "Online VS. Offline Coupon Redemption Behaviors", International Business & Economics Research Journal, Volume 9, Number 12, 2010, pp. 23-36.
- [3] T. Rueter (2011), Daily deal revenue will increase 138% this year, Local Offer Network says. [<http://www.internetretailer.com/2011/03/24/daily-deal-revenue-will-increase-138-year> (visited on 7 April 2013)].
- [4] NCH Marketing Services (2010), 2010 Coupon Facts.
- [5] V. Shankar, T. O'Driscoll, and D. Reibstein, "Rational exuberance: The wireless industry's killer 'B'", Strategy + Business, 31, 2003, pp. 68-77.
- [6] S. Srinivasan and K. Bawa, "Category-Specific Coupon Proneness: The Impact of Individual Characteristics and Category-Specific Variables," Journal of Retailing, 81 (3), 2005, pp. 205-14.
- [7] V. Kumar and B. Rajan, "Social coupons as a marketing strategy: a multifaceted perspective", Journal of the Academy of Marketing Science, 40, 2012, pp. 120-136.
- [8] M. Raskino, "Mobile coupons will reach right into your pocket", Gartner Group Research Note, 2001.
- [9] The Current State & Promise of Mobile Couponing - Mobile Marketing Update. Mobile Marketing Association (MMA), January 2013.
- [10] MMA [<http://www.mmaglobal.com/> (visited on 16-03-2013)].
- [11] GS1 - The global language of business. [<http://www.gs1.org/> (visited on 20-03-2013)].
- [12] Mobile Commerce: opportunities and challenges. A GS1 Mobile Com White Paper, GS1, February 2008 Edition.
- [13] Manual do Utilizador EAN•UCC. CODIPOR – GS1 Portugal.
- [14] Digital Coupon Management Standard Specification. GS1, June 2012.
- [15] Business Requirements Analysis Document (BRAD) - Digital Coupon Management. MSWG: B2C Digital Coupons, GS1, Draft 1.0, 1-Mar-2012.
- [16] Synchronising Data: Proven Benefits for Your Company. GS1 Brochure.
- [17] Validation and Electronic Clearing of Paper Coupons at POS Solves the Problems of Mis/Malredemption and Fraud for the Coupon Industry. ICN White Paper, Julho 2012.
- [18] The GS1 System Architecture. GS1, Issue 1.0, 14-February-2012.
- [19] GS1 Standards Knowledge Centre. [<http://www.gs1.org/gsm/kc> (visited on 25-03-2013)].
- [20] PacSis – Sistemas de Promoção e Marketing. [<http://www.pacsis.pt/> (visited on 20-03-2013)].
- [21] Coupons.com Incorporated. [<http://www.couponsinc.com/> (visited on 20-03-2013)].
- [22] Dico(re)2s [<http://www.dicore2s.com/> (visited on 04/04/2013)].
- [23] J. Sánchez-Silos, F. Velasco-Arjona, I. Ruiz, and M. Gómez-Nieto, "An NFC-Based Solution for Discount and Loyalty Mobile Coupons", 4th International Workshop on Near Field Communication, IEEE, 2012, pp. 45-50.
- [24] F. Borrego-Jaraba, P. Garrido, G. García, I. Ruiz and M. Gómez-Nieto "A Ubiquitous NFC Solution for the Development of Tailored Marketing Strategies Based on Discount Vouchers and Loyalty Cards", Sensors, 13(5), 2013, pp. 6334-6354.
- [25] J. Kim, Y. Lee, E. Kim, S. Kim, and M. Jung, "A Study of Coupons issuance System Considering of User Convenience Based on NFC", 3rd International Conference on Computer Science and Information Technology (ICCSIT'2013), Bali, Indonesia, 2013, pp. 10-13.
- [26] G. Van Damme, K. Wouters, H. Karahan and Bart Preneel, "Offline NFC payments with electronic vouchers", Proceeding, MobiHeld '09 Proceedings of the 1st ACM workshop on Networking, systems, and applications for mobile handhelds, 2009, pp. 25-30.
- [27] C. Chang and C. Sun, "A Secure and Efficient Authentication Scheme for E-coupon Systems", Wireless Personal Communications, Volume 77, Issue 4, 2014, pp. 2981-2996.
- [28] M. Aigner, S. Dominikus and M. Feldhofer, "A System of Secure Virtual Coupons Using NFC Technology" Proceedings of the Fifth Annual IEEE International Conference on Pervasive Computing and Communications Workshops (PerComW'07), 2007.
- [29] V. Patil and R. Shyamasundar, "e-coupons: An Efficient, Secure and Delegable Micro-Payment System", Information Systems Frontiers, Volume 7, Issue 4-5, 2005, pp. 371-389.
- [30] Forrester Consulting, "The State Of Digital Coupons, How Digital Coupons Are Adapting To Mobile And Omnichannel", Commissioned By RetailMeNot, 2014.

- [https://s3-prd-uss-rmn-corp-site ms.s3.amazonaws.com/filer_public/2d/9f/2d9f45d9-1508-45dd-9a12-8946d7793cf2/digital.pdf (visited on 08-01-2015)].
- [31] H. Wilcox, "White Paper - Mobilising your coupon offers", Juniper Research, 2008.
[http://www.juniperresearch.com/shop/products/whitepaper/pdf/MCS08_Whitepaper.pdf (visited on 08-01-2015)].
- [32] B. McBeath, "Mobile Coupons - The Missing Piece", ChainLink Research, 2012.
[http://www.clresearch.com/research/detail.cfm?guid=65A9C4B6-3048-79ED-9994-223F43CFC435 (visited on 08-01-2015)].
- [33] B. Farb, "GEOLOCATION: Advertising's Future OR 1984 Revisited?", CRM Magazine, Vol. 15 Issue 6, 2011, pp. 32-35.
- [34] K. Kang, K. Altinkemer, "E-coupons Strategy Problems in Location Based Advertisement", working paper, 2008.
[http://www.krannert.purdue.edu/academics/mis/workshop/papers/kanngPaper.pdf, (visited on 11-01-2015)].
- [35] Digital coupon clearinghouse. Patent publication number US20120284107 A1, Nov 8, 2012.
[http://www.google.com/patents/US20120284107 (visited on 20-03-2013)].
- [36] Digital Coupon System. Patent publication number US20120136712 A1, May 31, 2012. [http://www.google.com/patents/US20120136712 (visited on 20-03-2013)].
- [37] Method for Distributing and Redeeming Digital Coupons. Patent publication number US20120303440 A1, Nov 29, 2012. [http://www.google.com/patents/US20120303440 (visited on 20-03-2013)].
- [38] Selective Communication for Digital Coupon Distribution. Patent publication number US20120289255 A1, Nov 15, 2012. [http://www.google.com/patents/US20120289255 (visited on 20-03-2013)].
- [39] Method and apparatus for processing coupons/purchases based on radio frequency memory tag detection. Patent publication number US20120310720 A1, Dec 6, 2012.
[http://www.google.com/patents/US20120310720 (visited on 20-03-2013)].
- [40] Identifier-Based Coupon Distribution. Patent publication number US20120209686 A1, Aug 16, 2012.
[http://www.google.com/patents/US20120209686 (visited on 20-03-2013)].
- [41] Integrated coupon storage, discovery, and redemption system. Patent publication number US20120323664 A1, Dec 20, 2012. [http://www.google.com/patents/US20120323664 (visited on 20-03-2013)].
- [42] GS1 Position Paper on Barcodes for Mobile Applications. GS1, February 2012.
- [43] System and method for creating coupon offers and barcodes. Patent publication number EP2422307 A1, Feb 29, 2012. [http://www.google.com/patents/EP2422307A1 (visited on 20-03-2013)].
- [44] Couponing in the Digital Age: A Playbook for CPG Brands. [http://blog.360i.com (visited on 20-03-2013)].
- [45] Variable value coupons. Patent publication number US20120310719 A1, Dec 6, 2012. [http://www.google.com/patents/US20120310719 (visited on 20-03-2013)].
- [46] Third-quarter 2012 coupon distribution and redemption stats [http://www.jillcataldo.com/node/23791/ (visited on 10-05-2013)].
- [47] Mobile Couponing Trends in 2013
[http://www.weevermedia.com/mobile-marketing/mobile-couponing-trends-2013/ (visited on 10-05-2013)]
- [48] Coupon Trends Report – Insights and Analysis, 2012.
- [49] 5 Trends in Coupon Marketing for 2013
[http://smallbiztrends.com/2012/12/trends-coupon-marketing-2013.html/ (visited on 10-05-2013)].

Knowledge Intensive Evolutionary Algorithms

François de Bertrand de Beuvron*, Carlos Catania* and Cecilia Zanni-Merk*

ICube laboratory, BFO team,

INSA de Strasbourg, CNRS

France

Email: {debeuvron, catania, merk}@unistra.fr

Abstract—In this paper, we show through the resolution of a real problem, how knowledge engineering techniques can be used to guide the definition of Evolutionary Algorithms (EA) for problems involving a large amount of structured data. Evolutionary Algorithms have proven to be very effective in optimizing intractable problems in many areas. Various representations of the fitness functions (multi-objective EA), the genome and mutation / crossover operators adapted to different types of problems (routing, scheduling, etc.) have been proposed in the literature. However, real problems including specific constraints (legal restrictions, specific usages, etc.) are often overlooked by the proposed generic models. To ensure that these constraints are effectively taken into account, we propose a methodology based on the structuring of the conceptual model underlying the problem, creating a domain ontology suitable for optimization by EA. The real-world example, that is detailed throughout the article, belongs to the general field of medical assistance. The project focuses on the logistics involved in the transportation of the patients. Although this problem is a specific case of the heavily studied family of Vehicle Routing Problems (VRP), its specificity comes from the amount of data and constraints: in addition to costs, many legal or health considerations must be taken into account. Our approach is based on the development of a multi-objective genetic algorithm, which has to come up with the best itinerary taking all these constraints into account. We will show that a precise definition of the knowledge model with a domain ontology can be used to describe the chromosome, the evaluation functions, the crossover and mutation operators.

Keywords—*Knowledge engineering, multi-objective optimization problems, evolutionary algorithms*

I. INTRODUCTION

In this article, we will show that the use of knowledge engineering can greatly enhance the definition of an evolutionary algorithm for a real case. This study is the result of a collaboration between our team and an Alsatian SME that provides real data to deal with.

The project is developed under a healthcare system environment, specifically oriented to the transportation of patients, normally from or to some healthcare centre. The needs for developing an application arrives because of the fact that the enterprises, which take care of the logistics of the journeys have to manage a big amount of requirements and constrains at the moment of making an itinerary. This logistics affects many enterprise resources, like the employees and vehicles, which should be assigned in an efficient way in order to guarantee, among other things, the conformity of the patient and the satisfaction of certain law regulations.

The problem consists on satisfying the daily requests of the patients minimizing the costs and fulfilling certain constrains. The requests are basically for pick-ups and/or deliveries of the patients to or from their house to some healthcare centre.

There are different types of vehicles that can accomplish a journey and each of them has an associated cost. There are also the costs of affecting a crew, meaning a set of one or two employees, to a certain vehicle or to a certain patient.

Many studies concern the Vehicle Routing Problem (VRP) [1]. This large number of studies place themselves along two axes:

- the solving approach; mainly exact algorithms or meta-heuristics (stochastic or nature inspired algorithms) [2].
- the variants of the problem; including time windows constraints [3] or multiple heterogeneous vehicles with pickup and delivery [4], among others.

We have chosen to use Evolutionary Algorithms (EAs) for the specific problem to solve, since we will show in this article that some families of EA are particularly suitable for a knowledge driven definition.

Even if this problem belongs to the family of Vehicle Routing Problems, it soon became clear that the solutions proposed in the literature, as specific as they are [5][6][7], were not intended to take into account all the specific constraints of the problem. In addition, we believe that this situation is found in many optimization problems when the parameters are numerous and varied in nature. Thus, in our example, legal constraints such as the working time of ambulance attendants, or medical constraints such as the disinfection of the vehicles, or the personnel qualification, cannot be overlooked while minimizing costs or distances.

The entire legacy software environment, in particular the conceptual model of the information system, represents a body of knowledge and skills within the company, which should be used in the implementation of the optimization module. But strangely enough, we did not find in the literature any framework taking all of these needs into account. In fact, in the early days of genetic algorithms, much emphasis was put in the domain-independent nature of the basic crossover and mutation algorithms, working on a standard binary coded, fixed length chromosome [8]. The foundation of genetic algorithms relies on the exploitation of similarities (building blocks or schemas) in the chromosome. Using a specific chromosome representation and operators that emphasize meaningful building blocks in the application domain should improve the efficiency, as highlighted by C. Janikow in [9]. This general idea has led to numerous specific genetic algorithms based on domain knowledge. For example, in our problem, the distribution of tasks between ambulances can be seen as a special case of a set partition problem. This kind of problem led to the definition of specialized so-called Grouping Genetic Algorithms, and it

is no surprising that some further specialization has produced very efficient genetic algorithms for the pick-up and delivery problem [10]. In addition to the coding of the chromosome, knowledge may be used in other parts of a genetic algorithm: in the fitness function or the initial population, or to keep collective knowledge among individuals (Cultural Algorithms). A survey about such use of knowledge for the development of highly specialized EAs can be found in [11].

Our goal is different: we want to use domain knowledge to assist in the definition of the algorithm. That is why we propose a methodology centred on an extended domain ontology for the definition of evolutionary algorithms whose parameters and constraints represent a huge volume of structured data (this is what we call *Knowledge-Intensive Evolutionary Algorithms*).

The rest of the paper is structured as follows. In Section II, we present the main steps of the methodology we propose. In Section III, we detail the way of building an extended conceptual model in order to link the domain ontology with the EA own constructions. We then see how this structuring of the conceptual model can be used to define the evaluation functions (Section IV) then the chromosome, and associated mutation/crossover operators (Section V), ensuring that all the specific constraints of the problem are taken into account. Finally, we present in Section VI the preliminary results on the ambulance routing problem, before giving some conclusions and perspectives.

II. METHODOLOGY OVERVIEW

We propose a methodology in three phases (Figure 1).

A. Analysis

This phase consists in defining precisely the function to be optimized through a specific labelling of the domain ontology. The project objectives are identified by a team of domain experts, and a comprehensive list of constraints and costs related to the realization of these objectives is identified. A team of domain experts, incorporating expertise in Knowledge Engineering and in Evolutionary Algorithms is then formed. The aim is to link the goals / costs / constraints in the project with the generic concepts of EAs (evaluation functions, representation of the genome, mutation, and crossover) through a specific labelling of the domain entities. Four general, overlapping categories of domain entities are defined:

- 1) *Entities to Optimize*: entities whose values are to be determined by the optimization algorithm (e.g., which ambulance will take care of each client). The structure of the chromosome will largely depend on these entities and their relationships. It prefigures the output data structure of the evolutionary algorithm.
- 2) *Parameter Entities*: entities whose values act as costs or constraints (e.g., cost/km for a vehicle, legal maximum number of daily driving for an ambulance attendant). The calculation of the fitness functions depends on these entities. They prefigure the structure of the input data.
- 3) *Evaluation Entities*: entities whose values are objectives to optimize (e.g., minimize the total number of kilometres travelled by all ambulances or maximize the benefits). It must be ensured that each of these goals is represented by one or more fitness functions in the evolutionary algorithm.

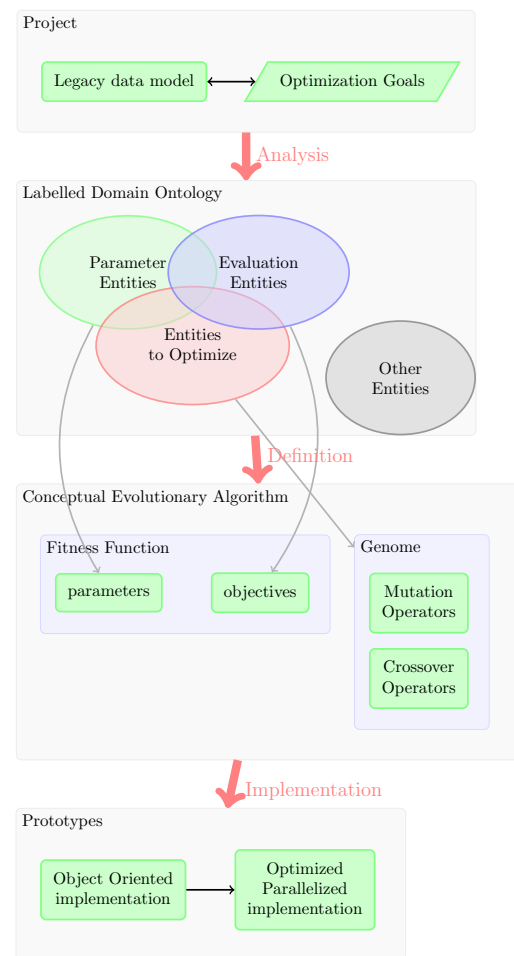


Figure 1. Methodology overview

- 4) *Other Entities*: entities that are not involved in the optimization (e.g., history of customer calls to the customer call centre).

The first three categories overlap. In fact, the membership of an entity to a category depends mainly on the usage of the values of its properties in the optimization problem. Some of the entities in the application domain may be “heterogeneous” when not all their properties serve the same purpose.

In the frequent case where the function to be optimized must be integrated into a larger existing system, most entities already exist in its data model. Most often, however, new entities will have to be introduced to have a more fine-grained representation of the entities to optimize.

We describe in Section 3 a specific labelling of the domain ontology to specify these categories.

B. Definition

From the labelled domain ontology defined in the analysis phase, the fitness functions, the structure of the chromosome and the associated evolutionary operators must now be defined. This process will be described in Section 4 for the fitness functions and in Section 5 for the genome and operators of the genetic operators.

C. Implementation

A first prototype using the EASEA platform [12], [13] was developed to test the feasibility of the approach. However, it soon became clear that it will not be sufficient to solve real problems. Limitations come from:

- the total memory footprint ($genomeSize \times populationSize$) is too large to be handled by a single computer
- the convergence time to an acceptable solution is too long.

For our application, the convergence time is critical: ambulance itineraries cannot be fully calculated off-line, because waiting times and consultation durations are largely unpredictable, and new journeys may be requested at any time, some of them urgent. The system must be able to take these changes into account and propose a new schedule in about one minute. Only a massively parallel implementation as proposed also by the EASEA platform can achieve such an efficiency.

Section 6 presents the evaluation of the first prototype. Notice that the highly specialized technical details concerning the parallel implementation are not discussed in this article.

III. STRUCTURING KNOWLEDGE FOR EVOLUTIONARY ALGORITHMS

The data model needs to deal with a huge quantity of information, taking into account all the constraints. Given the complexity of the data model, we have decided to formalize it as a domain ontology. The ontology will guide the definition of the evolutionary algorithm thanks to the formal relationships that appear among the entities in it.

An ontology is a knowledge representation of a domain or a field that provides conceptual resources for knowledge-based systems (KBS). It gathers and defines the set of objects that are known as belonging to the domain [14].

In general, an ontology is composed of entities sometimes called *concepts* or *classes* and relationships between these entities usually called *roles*, *properties*, or *attributes* if they are mono-valued. Within this paper we used *entity* and *property*. In fact, ontologies provide the conceptual and notional resources needed for knowledge formulation and for making knowledge explicit. Our domain ontology formalizes the main concepts concerning our problem, such as vehicles, crews, patients, addresses, journeys and so on (see entities below *DomainEntity* in Figure 2).

One of the essential components of the routing problem is *PlannedElement*. It is an event to be held at a given location (*Address*), which should start at a desired time (*requestedDate*), and should last a known or estimated time (*duration*). There are many subtypes of *PlannedElement*, some of them are shown in Figure 2:

- *PlannedEmployeeElement*: events related to an employee. For example, such employee shall be home by noon. On the same principle, there are also *PlannedVehicleElement* events related to a vehicle: the vehicle must be revised or disinfected, etc.
- *PlannedFleetElement*: an event that is not directly dependent on a vehicle or on a particular employee. The optimization algorithm will have to determine

which vehicle driven by which employees (*Crew*) will be assigned to the event. There are again two subtypes of *PlannedFleetElement*:

- *BusinessFleetElement*: this is the most classical event: a patient must be collected or deposited somewhere. These events are paired within a *Journey*, which includes the collection, the ride and the deposit of the patient. Of course, a patient who was picked up by an ambulance must be deposited by the same one. The optimization will obviously have to take this basic constraint into account. The *Journey* is associated to *PlannedElement* and not directly to a *BusinessFleetElement* since constraints related to the vehicle may be associated with *Journey*: for example, for some infectious patients, the ambulance must be disinfected immediately after the ride.
- *InternalFleetElement*: these are constraints that are not directly related to the patients (e.g., fetching a document in a hospital).

The assignment of the events to individual vehicles is the main goal of the optimization. A *PlanningLine* represents the list, ordered by increasing time, of the events supported by a given vehicle. A complete *Planning* is simply the set of the *PlanningLine* for all the vehicles.

To optimize itineraries, notions of distance and travel time are crucial. They are represented by the *Distances* and *Time for the Next (TFN)* entities. These data are huge: the ambulances of a large company may have to visit several thousands of addresses per day. As evolutionary algorithms may randomly test any path, the distance between any two addresses must be known or estimated. This is even more critical for the travel time, which usually depends on the time of day (see Section IV for more details).

Finally, the algorithm must take into account many additional constraints. For the sake of simplicity, only two appear in Figure 2:

- *CostCustomerVehicule*: vehicles are more or less suitable for patients. A patient in a wheelchair is easier to take care of (soft constraint) in a adapted vehicle. A patient who must travel lying down requires (hard constraint) a real ambulance.
- *CostCustomerEmployee*: common language, personal preferences, among others.

The ontology for Evolutionary Algorithms (EA) ontology provides generic, domain independent, entities and properties allowing to define what use will be made of entities and properties of the domain in the genetic algorithm. The EA ontology is therefore a characterization of the entities and properties of the domain ontology seen as individuals. A precise modelling would require for the EA ontology to be defined at a meta-model level over the domain ontology. The separation of entities into four categories as proposed in Section II should be represented as specializations of the meta-entity "Entity" itself. Similarly, a parameter used in the fitness function should be explicitly linked to some entities or properties in the application domain. For example, in the application domain of ambulances, the specific evaluation function "minimizing travel distances" should be connected



Figure 2. extract of the domain and EA ontology

(among others) to the property "address" of the "patient" entity. Unfortunately, not many formalisms allow the representation of relationships between different semantic levels (model and meta-model). In the ontology description language OWL 2 [15] for example, the possibility for a single object to be seen both as an entity and as an individual is reserved to the OWL-full expressivity, which is undecidable, and although theoretical studies have been conducted [16], few tools or reasoners, if any, allow to take into account this kind of structures.

We therefore chose a simpler model for the current version of our methodology in OWL-DL:

- two generic entities, *DomainEntity* and *EAEntity* are defined. All the domain entities are defined as subclasses of *DomainEntity*
- we define a generic object property *EAProperty* and three mutually exclusive sub-Properties *EAEvaluationProperty*, *EAOptimizableProperty*, *EAParameterProperty*. All object properties of the domain ontology involved in the genetic algorithm must be subproperties of one of these three specific properties depending on whether they are involved for optimization, in the evaluation, or as a parameter.
- For the same purpose, but for atomic domain knowledge, we also define a generic data property *EADataProperty* and three mutually exclusive sub-dataproperties. The top level EA object and data properties, and their specialization for the ambulance

example can be seen in Figure 3.

- The entity *EAEvaluationEntity* is defined as a subclass of *EAEntity* having at least one evaluation property :

$EAEvaluationEntity \sqsubseteq EAEntity$

$EAEvaluationEntity \equiv$
 (*EAEvaluationProperty* some Thing) or
 (*EAEvaluationDataProperty* some (boolean or dateTime or integer or real or ...))

EAParameterEntity and *EAOptimizableEntity* are defined similarly.

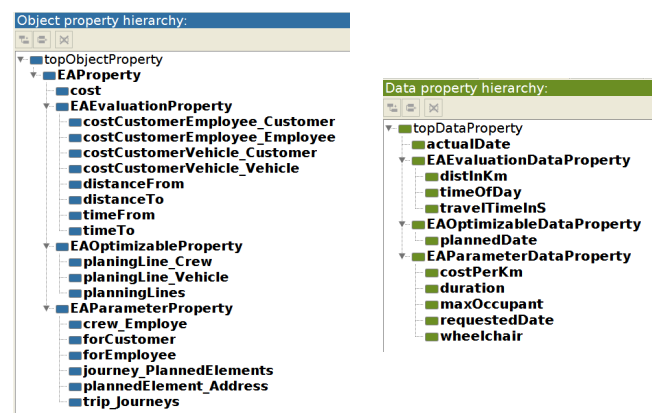


Figure 3. Object and data properties

The use of domain entities in the evolutionary algorithm is shown at two levels of granularity in the EA ontology: properties are classified according to their use within the EA, and every entity involved in the realization of the EA will be automatically classified into one or more sub-classes of *EAEntity* based on its properties. Thus, one can see in Figure 2 that *PlannedElement* plays a role both as a parameter and as an entity to optimize.

IV. MULTI-OBJECTIVE OPTIMIZATION AND EVALUATION FUNCTIONS

As stated in the previous sections, our problem is not a single-objective optimization problem, but a multi-objective one, because optimal decisions need to be made in the presence of trade-offs between two or more, eventually conflicting, objectives. Multi-objective approaches appear clearly as a possibility to solve our problem after a careful analysis of the constraints coming from the underlying data model. Moreover, beyond the classical advantages of these approaches, they are efficient in limiting a concentrated convergence of the solutions in a small subset of the Pareto front, which is very interesting for knowledge intensive evolutionary algorithms.

To be able to define a set of objectives, we first have to present a macro-objective and then break it down into a series of micro-objectives. The macro-objective is defined as: *Based on a set of required elements, vehicles, patients and employees, we have to generate a planning of itineraries that these vehicles will take so that the crews achieve the requested elements of the customers, minimizing the cost for the transport companies, maximizing service quality and observing all the restrictions that may appear in the context.*

Several micro-objectives have been deduced from this macro one, including:

- 1) Generate itineraries without delays or idle time between two different requests.
- 2) Generate itineraries that minimize the cost associated to the length of the trip.
- 3) Optimize the working time of the employees to avoid paying overtime or that they work less time that the legal number of hours per week.
- 4) Minimize the cost of the employees.
- 5) Minimize the cost of the vehicles.
- 6) Optimize the quality of service.
- 7) Balance the number of requests served by each vehicle.

These micro-objectives are, of course, in relation with the entities in the ontology in Figure 2.

For the two fitness functions detailed below, we will denote by V the set of vehicle, by PE the set of *PlannedElement*, and to each vehicle v_i , we associate the ordered list $P_i = [p_1^i, \dots, p_{n_i}^i]$, $p_j^i \in PE$ of *PlannedElement* assigned to the vehicle.

1) *Minimizing the cost associated with arrivals to a certain point in delay or in advance:* Each *Journey* between two *PlannedElements* p_j and p_{j+1} is represented in a temporal line (Figure 4) with three values, *RequestedDate* (RD), *Duration* (D) and the estimated time to arrive to the next point (TFN). The duration estimates the time needed in a point to take care of the patient. This time depends on multiple factors, for example, the size of the wheelchair if the patient needs one or the fact that the patient is in a stretcher or not.

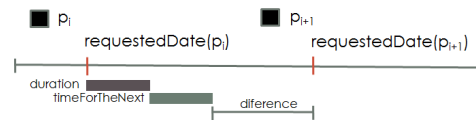


Figure 4. Temporal line between points p_j and p_{j+1}

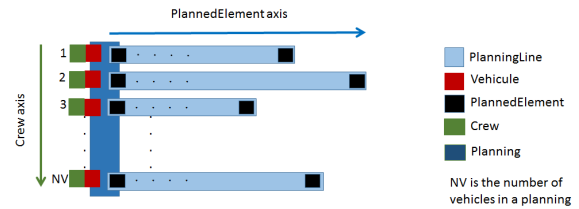


Figure 5. Structure of the chromosome

The TFN data are represented as a three-dimensional matrix (called *cubeTFN*), where each element represents the travel times between two addresses for a given (discrete) time in the day. Therefore, $cubeTFN_{ijt}$ is the time in seconds it takes to go from the point i to point j at time t .

The cost is represented through a piecewise function $f_0(x)$, where $x = RD_{j+1} - (RD_j + D_j + TFN_j)$ is the temporal difference shown in Figure 4. If this difference is negative (arrival in delay), the cost is quadratic; otherwise (arrival in advance), the cost is linear.

$$f_0(x) = \begin{cases} x^2 & , x < 0 \\ x & , x \geq 0 \end{cases}$$

Therefore, the objective can be formalized as:

$$\min \sum_{v \in V} \sum_{p \in P_i} f_0(RD_{j+1} - (RD_j + D_j + cubeTFN(p_j, p_{j+1}, RD_{j+1}))) \quad (1)$$

2) *Minimizing the cost associated to the length of the trip:* Ideally, the vehicles should attend points that are close to each other, in order to prevent the crews from driving long distances between two successive points in the itinerary. The cost to go from point p_j to point p_{j+1} is estimated from the distance matrix *cubeTFN* and the associated cost per kilometre of each *Vehicle* v_i . Therefore, if $x = (p_j, p_{j+1}, t)$ where t is the time of the day when the trip needs to be made, this objective can be formalized as:

$$\min \sum_{v \in V} \sum_{p \in P_i} costKm(v_i, cubeTFN(p_j, p_{j+1}, RD_{j+1})) \quad (2)$$

V. DEFINITION OF THE CHROMOSOME AND OF THE EVOLUTIONARY OPERATORS

When the genetic algorithm finishes its execution, it returns a planning based on the ontology, to which the structure of the chromosome is adapted (Figure 5). For the set of all the vehicles, the chromosome associates then a crew and a list of points (*PlannedElements*) that the vehicle needs to attend (this list of *PlannedElements* is the *PlannedLine* associated to each vehicle). All list of *PlannedElements* are sorted by

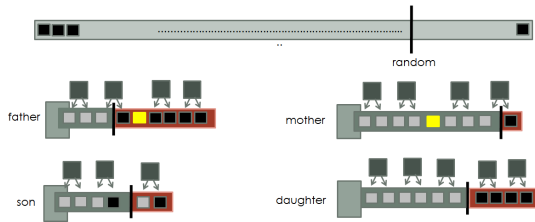


Figure 6. The *PlannedElementCrossover* operator

ascending requested Date (object property *requestedDate(p)* in model of Figure 3).

For the population to evolve, it is necessary to define a set of evolution operators. These should take into account all the aspects that are necessary for the planning to tend towards a possible final solution. There are two major axes that differ in the structure of a chromosome, the distribution of the *Crews* and the *PlannedElements* on the set of vehicles.

Making evolutionary changes in the two axes can affect all costs associated with the objectives, generating new populations with individuals of better quality. Several operators for the two axes have been defined, including crossover, mutation and swap. Only two of them are detailed below.

A. *PlannedElementCrossover*

As noted in Section III, *PlannedElements* may be re-grouped in a *Journey*. Each *PlannedElement* belongs to at most one *Journey*. We denote by $inJ(p)$ the set containing p and all the *PlannedElements* in the same *Journey* as p .

In Figure 6, the big long top rectangle represents the list of all the *PlannedElements* sorted by ascending time. To perform the crossover between two individuals, named mother and father in the figure, a *PlannedElement* p_r is randomly chosen in this list. For each *PlanningLine* i , we have the list P_i^f (resp. P_i^m) of *PlannedElements* affected to vehicle i in the father (resp. mother) individual. The set P_i^s of *PlannedElement* affected to vehicle i in the new son individual is defined by :

$$[l]P_i^s = \{p^f \in P_i^f; \exists p \in inJ(p^f) RD(p) \leq RD(p_r)\} \cup \{p^m \in P_i^m; \forall p \in inJ(p^m) RD(p) > RD(p_r)\}$$

Informally speaking, the son has the assignation of its father for early *PlannedElements*, and of its mother for later ones. We can also create a second derived individual (daughter in Figure 6) by reversing the role of the father and mother. The corresponding operations are schematically shown for one *PlanningLine* at the bottom of Figure 6.

B. *CrewMutation*

This operator is implemented by taking a random *Crew* of the list of all possible crews and make a simple swap between it and a randomly selected *Crew* in the individual, only if the number of employees is the same in both crews (Figure 7).

VI. EVALUATION

For a preliminary evaluation of the resulting algorithm, we consider the two evaluations functions described in Section IV. Evaluation is carried out under three different scenarios. In each scenario, we select an increasing number of *Vehicles*,

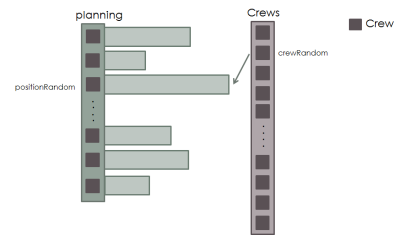


Figure 7. The *CrewMutation* operator

which have to satisfy the constrains of an also increasing number of *Journeys*. The input data of the algorithm consists of a number of *Journeys* randomly generated according previous information provided by a partner service medical company. As mentioned in Section III, each *Journey* includes constrains regarding the time, the cost, the number of occupants, etc.

All the three scenarios are evaluated using a population of 1000 individuals during 400 generations. The probability of the genetic operators are 0.8 for the crossover and 0.1 for the mutation.

Figure 8 shows the best and average fitness values for the three different scenarios along the different generations. The results considering the cost of the delay (equation (1)) are shown on the left side of the figure, whereas results for the cost for the length of the trip (equation (2)) are shown on the right. Notice that fitness values are shown in the y-axis while the generation number is shown in the x-axis.

As can be observed, for the three scenarios the algorithm has been capable of minimizing the values of both objective functions along the evolution process.

In the case of the delay function on the two first scenarios, we can observe that the improvement tends to slow down considerably beyond the generation 100. A different situation is shown on the third scenario, where the algorithm requires more generations to find a good solution. Clearly, as a consequence of having a larger problem, more evolution time seems to be required.

When considering the cost of the trip, the best values curve shows a more variable behaviour. However, even with this non monotonic behaviour, an improvement along the evolution process is still observed. In the second and third scenario the difference between average and best fitness values is bigger than the first scenario. Therefore, it could be possible that more evolution time be required.

Finally, we can observe that in those scenarios with the largest number of journeys to satisfy, average and best fitness values are also considerable larger. In particular, in the case of the cost fitness function we can observe best values starts at 70 (K€) for the third scenario, whereas best values start at 45 (K€) and 17 (K€) for the second and first scenarios, respectively. A similar pattern is also observed in the delay fitness function.

VII. CONCLUSIONS AND PERSPECTIVES

This article has shown, through the resolution of a real problem, how knowledge engineering techniques can be used to guide the definition of EA for problems involving a large amount of structured data.

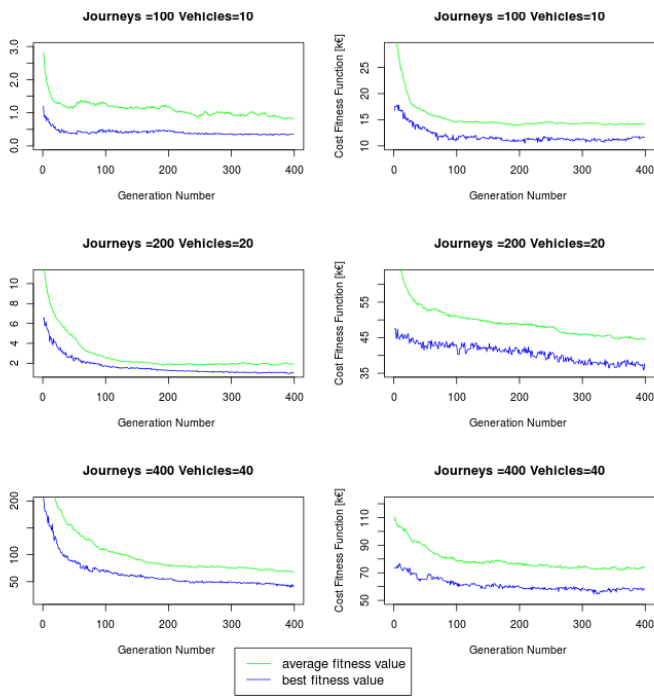


Figure 8. Fitness function values for different problem sizes

The first prototype, using the EASEA platform, has been evaluated. The idea behind this evaluation is just to demonstrate the feasibility of the methodology. Therefore, we have not fine tuned the EA parameters in order to get a solution suitable for the real-life scenario. Notice that such fine tuning will be carried out on a parallel version of the EA, which is currently being developed.

We also intend to use machine learning techniques in order to increase the efficiency of the EA. In fact, most of the *PlannedElements* are recurring in the same geographical area; the development of a module taking profit of the past experience is being undertaken to attempt guiding the population of the EA to more accurate and close-to-reality itineraries.

As evoked in Figure 4, each *PlannedElement* has an internal duration associated with it. Whether the point is a collection point or a deposit point, this internal duration depends specifically on the conditions in which the next ride needs to be made. This internal duration represents the amount of time taken for the patient to get in or out of the vehicle and depends on certain specificities of the patient, such as the need of a wheeling chair (that can have different sizes), or of a stretcher or of crutches, oxygen mask or perfusion. And, of course, this duration depends also on the age of the patient and on his general health state.

For the moment, some tests with linear regressions using WEKA [17] have been made, yielding encouraging results; although some occasional missing values in the input data (due mainly to oversights of the staff in the call centre that receives the requests) induces the need of using other techniques [18].

REFERENCES

[1] B. Golden, R. Raghavan, and E. Wasil, Eds., The Vehicle Routing Problem : Latest Advances and New Challenges, ser. Operations Research/Computer Science Interfaces Series,. Springer, 2008, vol. 43.

[2] S. J. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 3rd ed. Pearson Education, 2009.

[3] A. Garcia-Najera and J. A. Bullinaria, “An improved multi-objective evolutionary algorithm for the vehicle routing problem with time windows,” Computers & Operations Research, vol. 38, no. 1, 2011, pp. 287 – 300.

[4] A. S. Tasan and M. Gen, “A genetic algorithm based approach to vehicle routing problem with simultaneous pick-up and deliveries,” Computers & Industrial Engineering, vol. 62, no. 3, 2012, pp. 755 – 761.

[5] J. Berger and M. Barkaoui, “A parallel hybrid genetic algorithm for the vrp with time windows,” Computers & Operations Research, vol. 31, no. 12, 2004, pp. 2037 – 2053.

[6] O. Bräysy and M. Gendreau, “Vehicle routing problem with time windows, part ii: Metaheuristics,” Transportation Science, vol. 39, no. 1, Feb. 2005, pp. 119–139.

[7] N. Carrasquero and J. A. Moreno, “A new genetic operator for the travelling salesman problem,” in Proceedings of the 6th Ibero-American Conference on AI: Progress in Artificial Intelligence, ser. IBERAMIA ’98. London, UK, UK: Springer-Verlag, 1998, pp. 315–325.

[8] J. H. Holland, Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence. University of Michigan Press, 1975.

[9] C. Z. Janikow, “A knowledge-intensive genetic algorithm for supervised learning,” Machine Learning, vol. 13, no. 2-3, 1993, pp. 189–228.

[10] G. Pankratz, “A grouping genetic algorithm for the pickup and delivery problem with time windows,” OR Spectrum, vol. 27, no. 1, Jan 2005, pp. 21–41.

[11] R. Giraldez, J. Aguilar-Ruiz, and J. Riquelme, “Knowledge-based fast evaluation for evolutionary learning,” Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 35, no. 2, May 2005, pp. 254–261.

[12] P. Collet, E. Lutton, M. Schoenauer, and J. Louchet, “Take it EASEA,” in Parallel Problem Solving from Nature PPSN VI, ser. Lecture Notes in Computer Science, M. Schoenauer, K. Deb, G. Rudolph, X. Yao, E. Lutton, J. Merelo, and H.-P. Schwefel, Eds. Springer Berlin Heidelberg, 2000, vol. 1917, pp. 891–901.

[13] O. Maitre, F. Kruger, S. Querry, N. Lachiche, and P. Collet, “EASEA: specification and execution of evolutionary algorithms on GPGPU,” Soft Computing, vol. 16, no. 2, 2012, pp. 261–279.

[14] S. Staab and R. Studer, Eds., Handbook on Ontologies, ser. International Handbooks on Information Systems. Springer, 2004.

[15] B. C. Grau, I. Horrocks, B. Motik, B. Parsia, P. Patel-Schneider, and U. Sattler, “OWL 2: The next step for OWL,” Web Semantics: Science, Services and Agents on the World Wide Web, vol. 6, no. 4, 2008, pp. 309 – 322.

[16] B. Motik, “On the properties of metamodeling in owl,” J. Log. Comput., vol. 17, no. 4, 2007, pp. 617–637.

[17] I. H. Witten, E. Frankc, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques. 3rd Editio. Morgan Kaufmann Series in Data Management Systems, 2011.

[18] M. Saar-Tsechansky and F. Provost, “Handling missing values when applying classification models,” Journal of Machine Learning Research, vol. 8, 2007, pp. 1217–1250.

Guidelines for Social Media Mining for Innovation Purposes

Experiences and Recommendations from Literature and Practice

Robert Eckhoff, Mark Markus,
Markus Lassnig, and Sandra Schön
Innovation Lab
Salzburg Research Forschungsgesellschaft
Salzburg, Austria
markus.lassnig@salzburgresearch.at

Abstract— Social media in general and user-generated content are seen as potential sources for trend detection and innovation. Nevertheless, there are a quite variety of approaches and experiences. Within this contribution, we want to give an overview where we collect and discuss existing recommendations from literature and more. Building on lessons learned from literature and an expert discussion, we present guidelines for social media mining for innovation purposes. These guidelines especially build on experiences with an approach called innovation signals, which combines social media mining technology with an interpretative methodology. We explain guidelines, as “No tool will do your work automatically” or “Narrow your search”, and discuss our methodology of guidelines’ development as well as the results critically.

Keywords-innovation; social media monitoring; social media mining; guideline

I. INTRODUCTION

Social Media and its user-generated content is a valuable source for management tools and analysis that try to gain market and consumer insights. Weblogs, social networks or microblogging give insights in customers’ wishes, the image of a brand or new trends. Social media monitoring is also important to detect, select and analyze signs or facts for (future) innovation. In our contribution, we will give an overview about different approaches how social media mining is applied for innovation issues and provide insights into current research within the topic. The aim of our contribution is to deliver guidelines for researchers (and also practitioners) when they consider or plan to use social media mining for innovation purposes. Therefore, we will first introduce social media mining for innovation before we will present research question and research design as well as the guidelines we developed. Finally, we will discuss our results.

The developed guidelines should be interesting and helpful for consultants as well as researchers dealing with social media mining for innovation purposes.

II. SOCIAL MEDIA MINING FOR INNOVATION

A. Diverse uses of social media content and the idea of social media mining for innovation purposes

Social media are Web tools and services that allow to communicate, to collaborate, and to share. For example, social networks, discussion forums, Wikis, Weblogs or mailing lists are such applications. Within social media customers, colleagues, experts and others discuss brands, products and services, or related topics and issues. Therefore, social media is not only a way to share and discuss online, but also a good source of information for research and strategic planning. Marketing and PR departments of firms are interested in real-time monitoring of the Web to be able to react in time with a focus on customer satisfaction, customer relationship management, public relations or measurement of actions. Several tools help to monitor interactions within the web and send alerts for special keywords (e.g., the brand’s name) [1]. Sometimes active interaction within the social web [2] is called “social media listening”. In contrast to the in time monitoring of the social web the usage of social web content for mining purposes uses social media as source to find developments, new topics or interesting discussions. This approach is not only used for innovation purposes, but a quite common idea within innovation research [3]: Social web content is a cheap, non-reactive and authentic; therefore, it is a broad source for diverse methodologies and approaches to get insights or ideas for ongoing trends and future developments.

B. Approaches and Tools

There are several ideas what might be found related to innovation by social media mining [4]: Signs for potential innovations can be diffuse, but “weak signals” might have the potential of future impact on innovation [5]. For example Twitter messages are used as a source to detect weak signals for events [6]. Also, comprehensive overviews for tools that might be used for such weak signals detection [3] or social media monitoring in general [7] are available. The purpose of social media mining might also be the detection of ongoing developments or existing open innovations - These are innovations developed outside the enterprise, for example by customers [8] [9]. Mining for innovation also includes the detection and clarification of new trends, for example the

speed of adaptation of a new term or idea in a community. Some tools that are used to monitor or mine social media for innovation issues are for example, Attensity360, Brandwatch, Netbreeze Navigator, NM Incite - My BuzzMetrics, Radian6 or RapidSensitizer [3].

C. The Process of Social Media Mining for Innovation

Figure 1 gives an overview of the process of social media mining for innovation purposes. It builds on an overview of foresight with technologies [15]. We see it appropriate for a scheme of social media mining, but are skeptical about its linearity.

a) Preparation: This first stage includes all basic assumptions, researchers or customers make: What should be done? What is the goal of research?

b) Selection of tool(s): Another important issue is the usage of tools to support data collection, selection, and analysis.

c) Selection of sources and data: What concrete data should be used for analysis?.

d) Data collection: Data collection includes the sort of collected data, e.g., historical data, current postings, format of data.

e) Data processing: There are many ways for data processing. The selected tools and approaches influence the

processing (e.g., from keywords, detection abilities for sentiments).

f) Presentation of data: Finally, the results of the analysis should be presented in an appropriate and versatile way.

The last two phases of the process of innovation through social media mining (g) interpretation and discussion as well as (h) action are also influenced by, but not directly through, social media mining.

D. The Method of “Innovation Signals”

Exemplarily, we will now describe the research design of our project “innovation signals”. On the one hand, this is an example for a concrete research design [10] [11]. On the other hand, this is also a description of the background of the involved researchers and might explain their special experiences, even when we sum up and generalize the experiences and future guidelines.

The concept called “Innovation Signals” exploits user-generated content for strategic innovation purposes by combining quantitative and qualitative methods. The Innovation Signals research approach does not rely on technology alone, but unfolds in the development of social media mining technology in unique combination with an interpretative methodology. The process is described as following.

a) Set-up: The set-up of Innovation Signals research mimics the traditional research design of empirical social science. The main goal is to formulate research hypotheses and define conceptual search terms, which contain between 20 and 50 English and German keywords. Then, 40 to 50 publicly accessible social web sources (forums, communities, blogs, newsgroups) are identified and quickly assessed, according to a catalogue of criteria (e.g., quality of contents, length of contributions, intensity of contribution).

b) Detection and monitoring: The social media mining-based technology provides automatic detection of relevant keywords and topics of interest in sources selected before. It first extracts a large amount of user posts (e.g., 200,000 posts) and then, automatically detects emerging keywords, topics and sentiments from compiled discussions and user’s publicly available opinions.

The Innovation Signals technology provides answers to the questions in the context of product development and trend detection such as: How do users talk about existing products? What are critical issues? What issues are discussed very intensively? What are emerging topics? How do topics change over time? The technology enables experts to analyze and interpret detected innovation signals in an easy and intuitive way and also to save the most important posts for additional manual analysis and coding.

c) Identification and contextualisation of innovation signals: The automated analysis of textual content enables an efficient information processing, but the machine-processed information still remains ambiguous. In order to

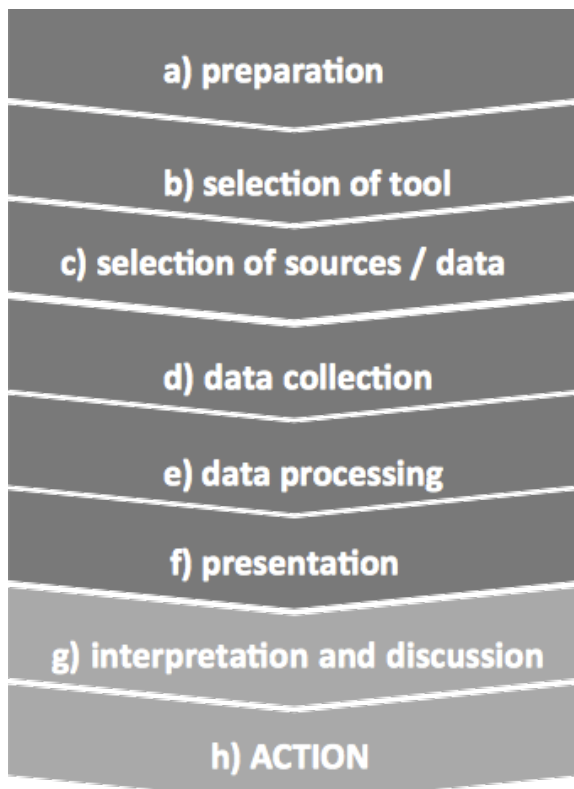


Figure 1. Process of social media mining for innovation purposes, based on [9], figure 3 (own variation)

enable effective research, the interactions in the social web must be structured additionally and analysed with social scientific methodology. This means to associate user generated content with relevant statistics, trends and theories to amplify the meaning of the information and to understand the consumers' conversations better and in a broader context.

d) Translation into business opportunities: This phase of the research process utilizes user generated content (in close co-operation with customers/companies) as an additional information source for strategic decision making with regard to the kind of innovation (product, process, business models, strategic innovation fields) to be pursued in order to determine the focus of the product innovation and market strategies and/or to detect new markets and new ideas.

The approach of innovation signals and the technology was developed and used within the project "Innovation Signals – Development of a Social Web Innovation Signals Amplifier System", funded by Austrian Research Promotion Agency. Three bigger and some smaller practical use cases have been delivered – for different branches and industry partners. The skiing industry, electric automobility, and the energy sector have been fields of application.

E. Quality criteria of Future Studies and Social Media Mining

According to our research, there is no comprehensive guideline available for the usage of social media mining for innovation purposes and future developments. Of course, there is a huge list of social media monitoring guides for general aspects as brand issues itself available, e.g., [12], [13]. An approach to develop guidelines is also to adapt quality criteria of future studies [14].

III. RESEARCH QUESTION

The guiding question of our research is: What comprehensive guidelines for usage of social media mining tools might be offered to researchers and practitioners in the field of innovation management and future studies?

IV. RESEARCH DESIGN

Within our study, we developed a multi-stage approach. First of all, we collected recommendations, lessons learned and comments from literature and the Web from (i) general social media monitoring guides and (ii) from usage of social media mining in research (literature). Short sentences and key advices were written on sheets of paper. Additionally, we made a list of quality criteria for future studies [15]. Building on this, concrete guidelines were to be developed in an expert workshop. The structured discussion included a systematic review of existing experiences as well as collection of our own expert assessments. In a final round, the main guidelines were approved.

V. RESULTS

The development process as well as the guidelines for social media mining for innovation purposes is presented in the following.

A. The development process

The expert workshop was held in September 2014. The moderator and three experts in the field are authors of this text. According to the process of social media mining for innovation from Figure 1 the recommendations and lessons learned were first of all completed with lessons learned from own work (see paragraphs about innovation signals approach).

The key source of the guidelines is the own practical experiences of the experts. They were asked to give advice to beginners, to think about pitfalls and also successes to collect a first set of issues for recommendations. Then, existing recommendations or criteria were used to systematically enrich and broaden the set of possible guidelines. Existing recommendations are sources from literature and the Web [12], [13], existing criteria for tools for special usage [3] as well as general criteria of future studies [14]. Another framework for orientation for the guidelines was the process of social media mining from Figure 1.

After collecting, a phase of re-arranging, combining and also selecting more important issues and topics started. When all experts were satisfied with the first sketch of guidelines, the authors formulated these. According to the formulation of the guidelines we did not find a guide for good guidelines, therefore we orientated our formulation of guidelines on good practice [16]. In a final phase, the recommendations were discussed by all experts and formulated as follows.

B. Guidelines

The following guidelines should support future social media mining, especially for beginners and also potential future customers.

a) No tool will do your work automatically: The expectation that a tool or set of tools might be able to do the work or the majority of work automatically is not realistic – not at all. In contrary, even sophisticated systems, or stand-alone tools are more or less only a small support for comprehensive social media mining; at least by now. Notably, a similar assessment is available for social media monitoring tools: "None of them [the tools] do what you want to do them" [17]. Even more concrete, the experts' experiences with sentiment analysis for innovation purposes or tool-delivered (automatically detected and defined) trends detection are unsatisfying by now.

b) Narrow your search. The more narrowly the search field or topic was defined, the merrier our results and customers' satisfaction was. This is of special importance when researchers are not already familiar with the sector of industry or topic. The data-driven approach is helpful, but only with a clear preference and focus in the beginning. For example, "car mobility" seems to be a good topic for mining

at a first glance. But thinking on the wide range of possible items, it might have been helpful to discuss in the very first meeting topics, which are really of interest, for example, “tuning of electric cars”. This opens the opportunity to mine for details within the concrete field of application.

c) The mining is manual social sciences; even better: The mining requires qualitative and quantitative social science. This is not only a consequence of the absence of useful stand-alone tools (see guideline a) but also building on the assessment of the power of the analytical approaches that are not easy to fulfill or overtake by algorithms. Manual work also includes that bottom-up (“data driven”) and top-down (“hypothesis driven”) approaches should be combined, as this is a typical approach of explorative social science. Re-adjusting working hypotheses, open minds and systematic procedures, e.g., coding and building of categories, are required working competencies for social media mining for innovation purposes. Some approaches directly build on social science experiences, for example, the so-called netnography approach [18]. The need of manual work by researchers is not only owed to missing possibilities for automation, but also the inevitable way to enhance results and realize smarter analyses. For example, the manual collection and selection of appropriate Web forums makes later analysis easier, as there is less “information noise” and unrelated content (eventually with same keywords) included.

d) The process is iterative rather than linear: Whereas typical illustrations of social media mining for innovation purposes are sketched as a linear process (see Figure 1), reality proves that successful social media mining for innovation has to be iterative and repetitive. A re-definition of goals, tuning of selected data and approaches, a combination of bottom-up (data-driven) and top-down (hypothesis driven) approaches, are needed and to be recommended in order to achieve satisfying results. Typically, an explorative process has no steady goals, hypotheses, sources, data, or possibilities to process.

e) Use well-selected topical discussion forums in the Web. The Web is full of opportunities and data where potentially interesting information might be delivered. The collection of data that is not directly about the key issues of the search requires further efforts while analyzing, especially if it contains similar key words and concepts in another context. Data extraction is simpler, if specialized forums are used. Besides this, we also try to choose such specialized forums for another reason. Topical discussion forums, e.g., on privately operated Wordpress-websites, more often not prohibit the analysis of public data (whereas most of the big commercial platforms, e.g., Facebook do).

f) Do not expect outstanding surprises when mining for innovation: According to several years researching for innovations for diverse customers, there had been no big surprises for insiders and experts within a certain domain. Presenting results from social media mining, even for vague

“weak signals” to insiders and experts, in no case delivered a real wow effect. The rate of completely new topics or ideas is probably very small, the pure information is seldom perceived (!) as surprising and new.

g) The role of researcher as the customers’ consultant is important: A solid know-how of social scientific methodologies seems to be a key factor, but this should be combined with a classical consultant expertise for a successful mining. A constant dialogue with the client, an effective management of user requirements and presentation skills is needed to be successful in form of clients’ satisfaction. Being a good consultant also includes being able to transfer issues into the client’s language.

h) Give results a meaning. Typically, social media mining delivers a long list of remarkable results, for example, a list of trendy topics or weak signals. Bringing social media mining to a success in form of customer satisfaction as well as potential action the results should get a meaning for the audience. There are several ways to give a good presentation, and not only a long list of themes and topics from your search field:

- It is not the list of topics, but the smart way to organize them in a model. This model might be the process of product development or service (e.g., from first users’ contact on).
- Time series might deliver a feeling for a development or trend.
- Numbers are helpful, e.g., how many posts are analyzed and how many of them concern a special topic?
- Exemplary statements (quotes) from the sources should be used to illustrate the numbers.
- Nevertheless, the presentation should not blur information of needs, information about possible solutions or prototypes.

i) Social media is not always a good or analysable source for innovation: Social media is in general a good source for many innovation purposes [e.g., 9]. Nevertheless, it is restricted: typically on languages (English, German, some others) and concerning the users (no offliners). Developing ideas of new mobile games should be done with Japanese data; and future developments of corrida (bullfight) should be done with Spanish content. Additionally, false friends or other meanings of words in another language complicate the analysis and should be considered while selecting sources or data. Whereas “e-bike” in German is a bicycle with an electronic motor (also called Pedelec), “bikes” for example, in English are also motorbikes. “Handy” is the German word for “mobile phones”, whereas “handy” is a common adjective in English.

j) Decisions within social media mining are a question of cost-value ratio, not of possibilities. Social

media mining is dealing with an endless source of data, especially if you broaden it from text to pictures and videos. As the mining itself is an explorative work or a spiral work (see d), the constant re-designs of the goal, source, collection and processing regularly needs a weighting up of workload and results and an assessment of the cost-benefit ratio.

C. Additional (concrete) recommendations

The experts also have two additional advices, non-related to the already mentioned guidelines.

- *Look for fitting thesauri!* Of course, thesauri are not available for any topic, but it is always worth to look if one is available. Using thesauri is a smart possibility to enhance results and analysis.
- *Use existing lists for social media mining and monitoring tools' evaluation!* Besides results of such evaluations and customers' feedback on the tools, the criteria used for evaluation might help you to get a clearer picture of your own needs. But always, take in mind guideline (a).

VI. CONCLUSION AND DISCUSSION

To sum up, the experts developed general advices as guidelines that may help to come to more realistic expectations concerning approaches and tools of social media mining for innovation purposes. Additionally, the guidelines give advice how the process might get better results and be more successful. The guidelines themselves are new and original, but in some points they are coincident with existing opinions and research [17]. As existing recommendations and lessons learned were part of the discussion, the presented guidelines can be seen as support for ongoing and future social media mining for innovation purposes.

Nevertheless, the development of guidelines and the guidelines themselves must be discussed critically. One point might be our set of experts. This might be considered as a limiting factor for the validity of the guidelines. The experts are only experienced in English and German speaking contexts and analyses, they have broad experiences with several tools and in several economic sectors, but of course there are still blind spots. As our current work with "innovation signals" focuses on technology-enhanced weak signals detection [19] this also might have influence on the results (compared with other social media mining for innovation approaches).

VII. NEXT STEPS

For future developments and potential adaptations of guidelines for social media mining, a broader base of experiences might be considered. Practically, there is no international network of researchers and practitioners that might be helpful for this special task. But social media mining for innovation purposes seem to be a developing [9], if not boosting approach, so future settings might allow broader developments of guidelines. So, we hope our

guidelines provide a valuable first step and a worthy support for beginners and practitioners alike.

Within our workshop and discussion several topics arouse that are still unclear from a research perspective, especially related to guideline (f) "Do not expect outstanding surprises when mining for innovation" [20]. Building on our experiences with social media mining, there seem to be space and possibilities to get better information and insights into innovation signals or trends. But the approaches, by now, are very limited concerning signals for disruptive innovations or other brand-new original issues that might become trendy in the future. Research that deals with historic data to evaluate existing approaches concerning their validity is also an interesting step forward.

ACKNOWLEDGMENT

Thanks to Kathrin Parson and Laurenz Giger for their support within the project.

REFERENCES

- [1] I. Stavrakantonakis, A. Gagiou, A., H. Kasper, I. Toma, I., and A. Thalhammer, A. "An approach for evaluation of social media monitoring tools" in: D. Fensel, H. Kett, and M. Grobelnik (Eds.), *Common Value Management*, Stuttgart Fraunhofer-Institut für Arbeitswirtschaft und Organisation, 2012, pp. 52-64.
- [2] R. W. Helms, E. Booiij, and M. R. Spruit, "Reaching out: Involving users in innovation tasks through social media", in: *ECIS 2012 Proceedings*, paper 193, Online: <http://aisel.aisnet.org/ecis2012/193> (2014-09-15).
- [3] K. Welz, L. Brecht, J. Kauffeldt, and D. Schallmo, "Weak signals detection: Criteria for social media monitoring tools", *Proceedings of the 5th ISPIM Innovation Symposium: "Stimulating Innovation: Challenges for Management, Science & Technology"*, 09-12 December, 2012, Seoul, Korea.
- [4] R. Zafarani, M. Abbasi, and H. Liu, *Social Media Mining. An Introduction*, 2014, Cambridge University Press.
- [5] I. Ansoff, "Managing Strategic Surprise by Response to Weak Signals", *California Management Review*, 18, 2, 1975, pp. 21-33.
- [6] B. Song, "Weak Signal Detection on Twitter Datasets. A non accumulated approach for non-famous events", Master Thesis TU Delft, 2012, Online available: <http://www.cs.uml.edu/~hachreka/files/related/Thesis.pdf> (2014-07-01)
- [7] I. Stavrakantonakis, A. Gagiou, A., H. Kasper, I. Toma, I., and A. Thalhammer, "An approach for evaluation of social media monitoring tools", in: D. Fensel, H. Kett, and M. Grobelnik (Eds.), *Common Value Management*, Stuttgart Fraunhofer-Institut für Arbeitswirtschaft und Organisation, 2012, pp. 52-64.
- [8] H. W. Chesbrough, *Open Innovation: The New Imperative for Creating and Profiting from Technology*, Harvard Business Press, 2006.
- [9] F. Piller, A. Vossen, and C. Ihl, "From Social Media to Social Product Development: The Impact of Social Media on Co-Creation of Innovation", 2012, Online: <http://www.alycante.it/wp/From-Social-Media-to-Social-Product-Development.pdf> (2014-09-15).
- [10] M. Markus, R. Eckhoff, and M. Lassnig, "Innovation Signals in Online-Communitys – ein komplementärer analytischer Ansatz", *HMD - Praxis der Wirtschaftsinformatik*, 293, 10/2013, pp. 13-21.
- [11] M. Lassnig, M. Markus, R. Eckhoff, and K. Wrussnig, "Prospects of technology-enhanced Social Media Analysis for Open Innovation in the Leisure Industries" in: R. Egger, I. Gula, and D. Walcher, Eds., *Open Tourism – Open Innovation, Crowdsourcing and Collaborative Consumption challenging the Tourism Industry*. Salzburg, 2013.

- [12] L. Bush and L. Glazier, „3 guidelines for using social media monitoring during a PR crisis“, Prdaily.com, Post from 2013-09-13, Online: http://www.prdaily.com/Main/Articles/3_guidelines_for_using_social_media_monitoring_dur_15189.aspx (2014-09-15).
- [13] Socialmedia hq (without year). “The Definitive Guide to Social Media Monitoring for Business”, Online: <http://www.socialmediahq.com/pdf/the-definitive-guide-to-social-media-monitoring-for-business.pdf> (2014-09-15).
- [14] M. J. Boon, E. Rusman, and M. R. Klink, “Developing a critical view on e-learning reports: Trend watching or trend setting?” *International Journal of Training and Development*, 2005, 9(3), pp. 1-27.
- [15] J. Keller and H. A. von der Gracht, “The influence of information and communication technology (ICT) on future foresight processes — Results from a Delphi survey”, *Technological Forecasting and Social Change*, Volume 85, June 2014, pp. 81-92.
- [16] Deutsche Forschungsgemeinschaft (DFG), *Safeguarding Good Scientific Practice*, 2014, Wiley. Online available: http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahme_n/download/empfehlung_wiss_praxis_1310.pdf (2014-09-15).
- [17] J. Falls, “Where Social Media Monitoring Services Fail”, published at socialmediaexplorer.com, April 1, 2010, Online: <http://www.socialmediaexplorer.com/social-media-monitoring/where-social-media-monitoring-services-fail/> (2014-09-15).
- [18] G. Jawecki and J. Fuller, “How to use the innovative potential of online communities? Netnography – an unobtrusive research method to absorb the knowledge and creativity of online communities”, *International Journal of Business Process Integration and Management*, 3, 4/2008, pp. 248-255.
- [19] R. Eckhoff, M. Markus, M. Lassnig, S. Schön, “Detecting Weak Signals with Technologies. Overview of current technology-enhanced approaches for the detection of weak signals”, *International Journal of Trends in Economics Management & Technology (IJTEMT)*, volume III issue V, October 2014, URL: http://www.ijtemt.org/vol3issue5/1_Detecting_Weak_Signals_with_Technologies_Vol_III_Issue_V.php
- [20] R. Eckhoff, M. Markus, M. Lassnig, S. Schön, “No outstanding surprises when using Social Media as source for weak signals? First attempt to discuss the impact of social media sources to detect surprising weak signals”, *Proceedings of The Ninth International Conference on Digital Society (ICDS) 2015*, in Lisbon, Portugal.

The Application of Machine Learning to Problems in Graph Drawing

A Literature Review

Raissa dos Santos Vieira
Hugo Alexandre Dantas do Nascimento
Wanderson Barcelos da Silva

Institute of Informatics
Federal University of Goiás
Goiânia - GO, Brazil

Email: {raissavieira, hadn, wandersonsilva}@inf.ufg.br

Abstract—Graph drawing, as a research field, is concerned with the visualization of information modeled in the form of graphs. The present paper is a literature review that identifies the state-of-the-art in applying machine learning techniques to problems in graph drawing. We focused on machine learning strategies that build up and represent knowledge about how to draw a graph. Surprisingly, only a few pieces of research can be found about this subject. We classified them in two main groups: the ones that extract knowledge from the user by human-computer interaction and those that are not based on data directly gathered from users. The study of these methods shows that there is still much to research and to develop regarding the application of machine learning to graph drawing. We suggest directions for future research on this area.

Keywords—Graph Drawing; Human-Computer Interaction; Machine Learning.

I. INTRODUCTION

Graphs are mathematical models defined as a set of vertices and a set of edges. They are widely used to represent physical and abstract entities and their relationships. Often, it is necessary to draw a graph, that is, to construct a geometric representation of its vertices and edges [1]. For this aim, it is common to choose a standard graph-related convention (for example, drawing vertices as circles and edges as straight lines) and a set of aesthetic criteria (such as displaying edges with uniform orientation, minimizing edge crossings and presenting symmetry).

When a graph contains only a few vertices and edges, it can easily be drawn manually. However, as the size of the graph increases, manual drawing becomes more difficult and time consuming. The most common strategy for drawing medium to large-size graphs – ranging from hundreds to thousands of vertices and edges – is through the usage of automated techniques that incorporate a set of aesthetic criteria and apply algorithms for finding aesthetically pleasing drawings. A number of computational systems for drawing graphs exist based on this approach. Including, we have GraphViz [2] and Gephi [3].

Drawing a graph by a computational process, on the other hand, also creates many difficulties. One of them is that the search for drawings of good quality drawings with several aesthetic criteria is an NP-Hard problem [1]. In addition, some aesthetic criteria are in conflict, so that the improvement of a

drawing in respect to one criterion may imply a reduction of other aesthetical aspects. Furthermore, the drawing of a graph is essentially a subjective task – some users may prefer to satisfy some particular aesthetic criteria, different from the preferences of other users.

For this reason, even with the use of heuristics, there is still a need for human intervention to assist in obtaining good quality graph drawings. This was perceived very early in the advent of the graph drawing research field, resulting in the inclusion of human-computer interactive resources in most graph drawing systems. Such resources help to tailor the drawing towards satisfying aesthetic criteria that are not fully treated computationally, and to escape from local minima, when the graph drawing process involves an optimization model.

In the current paper, we present a literature review of the use of machine learning techniques for graph drawing. This is a much more complex challenge than merely having a fully automatic graph drawing system or a system with a few simple interactive tools. We shall comment on computational approaches that attempt to acquire knowledge that can be used to help drawing graphs. As can be seen, there are few reports of research on this subject. However, some interesting ones have been found and their study may lead to promising lines of future research.

The remainder of the paper is organized as follows: in Section II, graph drawing definitions are presented; Section III provides an overview of the application of machine learning techniques to graph drawing; Section IV summarizes the characteristics of the existing approaches; finally, in Section V, we draw our conclusions and suggest future works.

II. GRAPH DRAWING

A *finite undirected graph* G is an ordered pair (V, E) of finite sets V and E , where V is a set of *vertices* representing a set of discrete objects, and E is a set of unordered pairs $\{x, y\}$ of distinct elements in V , termed *edges* between vertices x and y . In a *directed graph*, E is a set of ordered pairs of vertices such that an edge $e = (x, y)$ connects x to y , but the reverse is not true, unless there is an edge $e' = (y, x)$ in E . A *two-dimensional drawing* of a graph $G = (V, E)$ is a function that associates each vertex and edge of G with geometric objects in a drawing space.

The graph drawing process begins with a computational representation of a graph (e.g., an adjacency matrix or a vertex-edge incidence list) and the selection of a *graphical convention*. Vertices are usually depicted as circles or squares while the edges are commonly represented by polygonal lines or arcs. Then, a set of aesthetic criteria are chosen in order to determine the aesthetic quality of any drawing of the graph. Aesthetic criteria, frequently employed as soft constraints, are:

- Minimizing the number of edge crossings;
- Displaying symmetries;
- Distributing the vertices evenly in the drawing area;
- Showing all edges with uniform length; and
- Arranging the edges in the same direction as much as possible (in the case of directed graphs).

In addition to the aesthetic criteria, soft constraints, hard “drawing constraints” can be imposed. While an aesthetic criterion may be neglected to a certain degree, all the hard drawing constraints must always be satisfied. Examples of drawing constraints are:

- Avoiding vertex over placement and
- Requiring some vertices and edges to be located in given fixed positions.

In the graph drawing research area, the drawing problem is studied according to the class of graphs considered, e.g. undirected, directed, planar or tree graphs. For each particular class, there are graph conventions, aesthetic criteria and hard constraints, all of relative importance.

A drawing of a given graph that reflects these details can be attempted. With a large drawing area, usually different drawings can be generated for non-trivial graphs. Not all of these drawings are of practical interest. The most desirable ones are those that best satisfy the aesthetic criteria, since they have a higher chance of being readable, that is, helping understand the inherent graph structure.

One of the difficulties of obtaining useful drawings, however, is that finding the best drawing for many aesthetic criteria is a NP-Hard problem [1]. Also, some aesthetic criteria conflict with the other. In this case, there may be no drawing that optimally satisfies two or more adopted aesthetic criteria. In such situations, there must be a trade-off between the criteria.

The definition of what is a “good” drawing can also be a very challenging task. The opinions of which aesthetic criteria to adopt may vary significantly between users. Sometimes, a common set of aesthetic criteria is desirable by all users, but they vary in their relative importance. For instance, one user may prefer edge crossing minimization over drawing symmetry, while another user may prefer the opposite. Many computational approaches to evaluate the quality of a graph drawing employ a weighted function of the given aesthetic criteria with weights provided in advance by the user.

In some situations, the user’s preferences are imprecise and difficult to model mathematically. In this case, the users have to draw the graph manually, possibly in an incremental way until an acceptable quality of the drawing is reached.

Thus, we can conclude that human interaction is necessary in many current graph drawing systems because these systems do not automatically and adequately deal with all drawing issues that may be of concern to the users.

In a previous work, Nascimento and Eades [4] investigated interactive optimization systems for graph drawing. They also proposed a framework that combines graph drawing algorithms and human interaction in an optimization process. The main result of their study was that the combination of graph drawing algorithms with human interventions help to achieve better results than having the algorithms and the users working independently. Furthermore, the authors identified certain interactive actions already being performed by users that could be used as the basis for new graph drawing algorithms. They then suggested applying machine learning techniques to learn and automatize such actions.

The idea of using machine learning to graph drawing goes far beyond the traditional interactive graph-drawing approaches, as it tries to computationally empower existing algorithms while releasing users to carry out more suitable roles. The following section presents a review on existing approaches that, in some ways, pursue this idea.

III. USING MACHINE LEARNING FOR GRAPH DRAWING

We performed a search of the relevant scientific literature in order to identify existing graph drawing methods that use machine learning techniques. The main literature databases investigated were: the bibliographic database of the Institute of Electrical and Electronics Engineers (IEEE), the database of the Association for Computing Machinery (ACM), Scopus, CiteSeer, Science Direct and the Web of Science. Our search string combined the terms: Graph Drawing, Graph Layout, Machine Learning, Expert Systems, Task Learning, Case-Based, Knowledge-Based, User Preferences and Artificial Intelligence.

The search identified 86 papers. Four other references that we previously knew of were added. We then analyzed all these papers and reduced the bibliography to only 11, all of which actually use machine learning techniques to help to produce good drawings of graphs. This final set of papers was divided into two main groups, presented next:

- 1) *Approaches that learn from human interaction* – these are approaches that learn information from users based on their interactions to a graph drawing system. They are presented in Section III-A;
- 2) *Approaches that are not based on human interaction* – we consider here approaches that gather and evolve knowledge about how to draw a graph from the results of other automatic graph drawing algorithms or from the graph structure itself. These approaches are presented in Section III-B.

We describe other details about the approaches that allows a more refined classification, later in the paper. Some aspects that we analyze include: the goal of the learning process (defining the quality of the drawing, or improving the convergence of an optimization process towards a good drawing), the class of graphs being drawn, the type of information interactively provided by the user for the learning process, the learning method, and whether or not the acquired knowledge can be reused to draw other graphs.

A. Approaches based on human-computer interaction

Since the choice of aesthetic criteria and their importance is an inherently subjective task, Neto [5] and Neto and Eades [6]

developed an approach for learning users' preferences. They were in the first to propose the automatic learning of human knowledge for drawing graphs. For this, they created an interactive system in which the parameters of an objective function (inferring the desirable aesthetics) and the setup parameters of a simulated annealing graph drawing method are automatically adjusted. As shown in Figure 1, the system consists of: a graph drawing editor, for human operators to draw the graph according to their preferences; a learning module that, implicitly, observes the users' actions and tries to learn and reproduce their results; and a knowledge base consisting of weights for a set of predefined aesthetic criteria (these weights appear in the objective function) and some parameters for initializing the simulated annealing process. Using the editor, the user provides a drawing of the graph. While doing this, the system performs two tasks autonomously: (1) it measures the aesthetic criteria of the user-generated drawing (2) and it runs a process for automatically learning all parameters necessary for reproducing the drawing. The learning process is also based on a simulated annealing strategy. The knowledge acquired by the system may help to recreate a drawing for other very similar graphs. However, the authors mentioned that it is not clear whether the aesthetic parameters can be meaningful to draw other graphs.

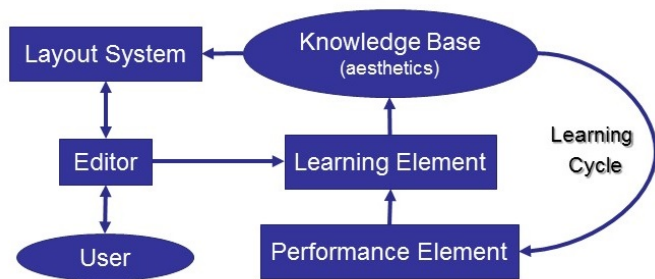


Figure 1. Learning model (adapted from [5]).

Some other approaches for the learning of aesthetic criteria are based on evolutionary computation techniques, such as genetic algorithms. This has been done by Masui [7], Rosete-Suárez *et al.* [8], Barbosa and Barreto [9], Bach *et al.* [10] and Spönemann [11]. In general, these authors present systems that collect information from human interaction that helps to clarify and to adjust a fitness function, as illustrated in Figure 2. Their goal is to refine the definition of the graph drawing problem (in the sense of identifying what kind of drawing addresses the subjective aesthetic criteria).

Following this line of thought, Masui [7] proposed having the user evaluating drawing examples, which are rated as “good” or “bad”. This evaluation is done in a preliminary stage. Figure 3 illustrates some examples given by the user to the system. Such pieces of information then serve as a reference for the algorithm to infer the desirable aesthetic criteria. The approach was developed only for directed graphs. After setting the fitness function based on examples given by the user, an evolutionary process is run in order to produce drawings of these graphs.

Rosete-Suárez *et al.* [8] built a system that also acquires user preferences, but from the manual scoring of drawing

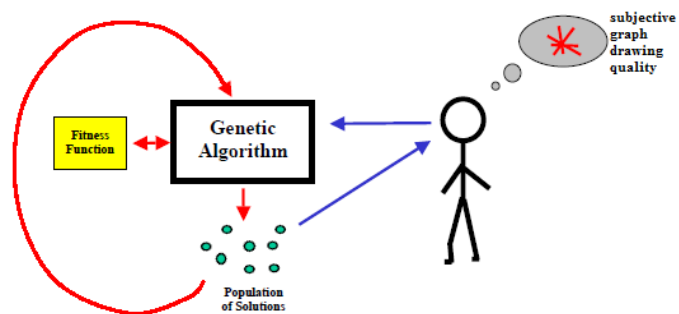


Figure 2. Interactive evolutionary computation: the user evaluates graph drawing solutions and helps to refine a fitness function (adapted from [4]).

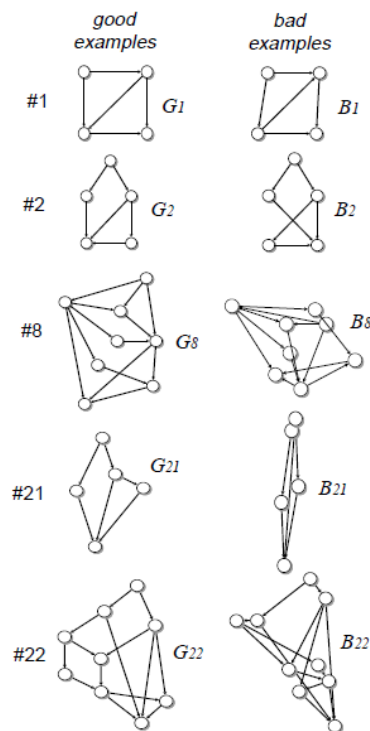


Figure 3. Good and bad example layout pairs given to the system reported in [7].

solutions in the population of a particular genetic algorithm. They worked with undirected graphs. When running the genetic algorithm, the system asks the user to give scores for six graph drawings from the current population. The learning of weights for a fitness function occurs from such an evaluation. The fitness function is adjusted with these weights to reflect the importance to the user of each aesthetic criterion. The interactive process also tries to improve the convergence of the genetic algorithm, since solutions with good aesthetic quality evolved more quickly than in some other evolutionary approaches without human interaction.

Likewise, Barbosa and Barreto [9] employed user evaluation to undirected graph drawings. However, they used a process called co-evolution, in which the set of weights of a fitness function is evolved indirectly using information from both the user and an internal evolutionary algorithm. Furthermore, the

user ranks the solutions in the current population instead of giving full scores. Figure 4 shows some drawings that were presented by their system to the user for evaluation.

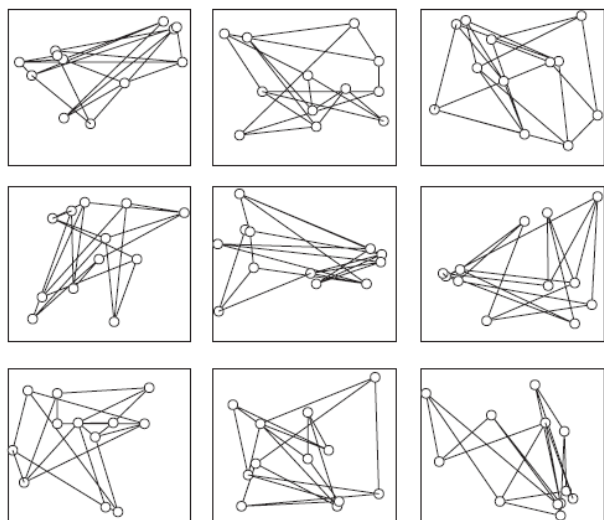


Figure 4. A sample from the initial population (from [9]).

Bach *et al.* [10] presented a system called “GraphCuisine”, a random drawing generator for undirected graphs with a populational viewpoint. A random initial population is generated and an evolutionary algorithm improves the quality of this population. Then the user selects solutions they deem to be the best by visual inspection. When multiple graphs are selected, the system tries to infer what graph characteristics are important to the user and adjusts the fitness function to represent them. This interactive approach aims to refine the problem of how to produce good graph drawings and to improve the convergence of the genetic algorithm. The system interface to select the best drawings is shown in Figure 5. Note that there are options for the users to modify the generation interval in which they want to view the population, the population size and how many drawings should be presented for selection.

Lastly, Spönemann [11] makes use of two interactive human inputs with an evolutionary approach: the users can directly manipulate the weights of a fitness function by changing visual sliders, or they can select good drawings from the current population. Both actions result in the automatic adjustment of the weights in the objective function, representing aesthetic criteria. Figure 6 presents the user interface showing drawings for individuals of the current population. There is a checkbox below each graph drawing, which enables users to select the drawing for the purpose of the adaption of the weights. The system works with both undirected and directed graphs.

An advantage of using genetic algorithms for interactive optimization is that it naturally produces several alternative solutions for user evaluation and it is flexible in dealing with different aesthetics criteria and constraints.

B. Approaches not based on human-computer interaction

Other approaches exist that use well-known machine learning techniques for drawing graphs, but that do not take into consideration data provided by human operators.

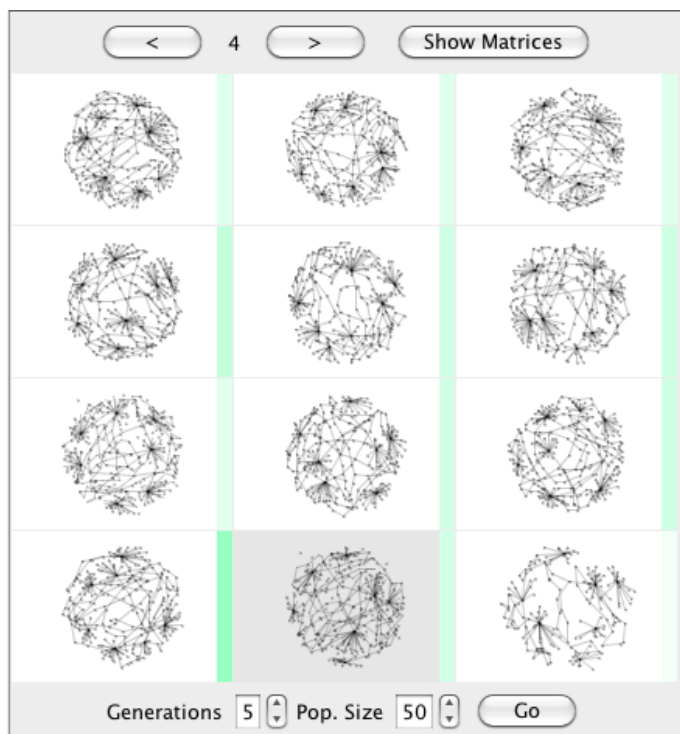


Figure 5. The population view, displaying representative graphs of the current population (adapted from [12]).

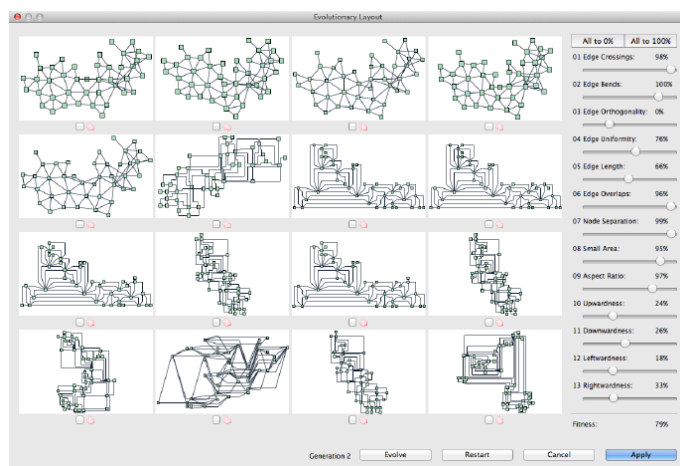


Figure 6. User interface for evolutionary meta layout (from [11]).

We start mentioning three reports of research that are based on neural networks.

Cimikowski and Shope [13] presented a parallel neural network algorithm that minimizes edge crossings in drawings of nonplanar graphs. They based their work on that of Takefuji and Lee [14] for computing a maximal graph planarization. The use of a neural-network representation for solving optimization problems was proposed by Hopfield and Tank [15]. After them, several developments followed.

Cimikowski and Shope dealt with graph drawings in the form of arc diagrams (or linear embeddings), as illustrated in Figure 7 for K_6 , the complete graph on six vertices. In

their graphical convention, the vertices are placed along a horizontal line and the edges are drawn as semicircles in one of the two half planes bounded by the line. Their optimization approach consists of modeling each edge of the graph as two neurons, representing the possibility of the edge being drawn above or below the line. Every neuron can be set to state 0 or 1, and this indicates distinct configurations for the position of the edges (above or below the line, not placed at all or in an inconsistent condition). An energy system is then created, containing excitatory and inhibitory forces that push the neurons to a state that minimizes the number of edge crossings while ensuring that all edges are placed. Through an iterative process, the system converges from a randomly-defined initial state to a stable configuration with the minimal number of edge crossings. In terms of the quality of the final results and processing time, the approach proved to be more competitive than a conventional greedy algorithm for the same problem.

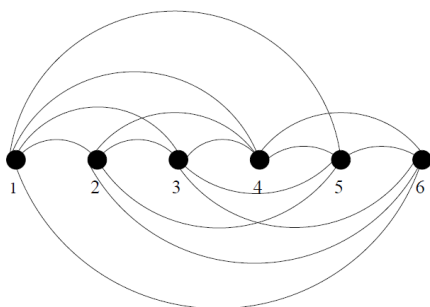


Figure 7. Linear embeddings of the complete graph K_6 with 3 crossings (from [13]).

Wang and Okazaki [16] improved the method proposed by Cimikowski and Shope [13]. They modified the internal dynamics of the neural network to permit temporary increases in the energy function in order to help the network to escape from local minima.

Finally, Meyer [17] presented an approach for graph drawing based on a competitive learning algorithm and on an extension of ideas used in unsupervised neural networks and self-organized maps. An association between a self-organized network and a graph drawing problem was established. The graph to be drawn is viewed as a neural network to be trained – every vertex of the graph is a neuron in the network; the position of a vertex v in the drawing is, in fact, described by the weight vector of the connections of v to other neurons. Drawing a graph is then treated as the problem of training its related network, that is, finding the proper sets of weights for the neurons. The training method consists of a competitive learning algorithm that updates the weights associated to a winner neuron v and to its neighbors so that they get closer to a random stimulus (to a point randomly chosen on a mesh over the drawing area). Ideally, the winner is the neuron whose weights are the closest to the stimulus. This process is iterated with new winners, with the changes being less significant at each time. The final result of the algorithm is a set of weights (coordinates) that visually highlights the structure of the graph. Figure 8 shows two drawings of K_6 , obtained using a triangular mesh and a rectangular mesh for the random

stimuli. The approach is also able to draw graphs in a three-dimensional space and to deal with some spatial aesthetics criteria.

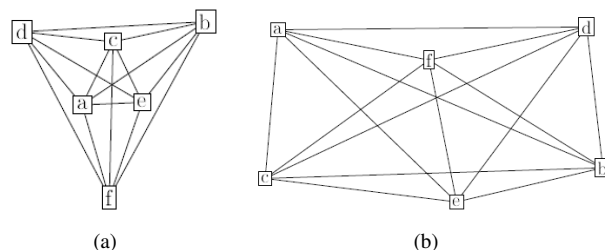


Figure 8. The complete graph K_6 : (a) triangular layout, (b) rectangular layout (from [17]).

Despite neural networks and self-organized maps being very catchy names, we must point that these three related approaches deviate from the machine learning nature in which we are interested. They really do not build a general neural network that learns aesthetic criteria or hints about how to draw a graph. What their methods do is to model the graph structure and the drawing problem as a network coupled with an energy system. The methods then improve the state of the network in order to minimize the system energy, usually resulting in a good quality graph drawing. Unfortunately, there is not resultant learned knowledge that can be reused, and every new graph has to be drawn by repeating the whole optimization process from scratch.

A different and more interesting learning approach is the one proposed by Stolfi *et al.* [18], using asynchronous teams (A-teams) for drawing graphs. By experimenting with teams of graph drawing heuristics, they verified that some sequences of heuristics (alternating over a drawing of a graph) could produce better results than other sequences. They then termed such good sequences as “pedigrees” and built a system for finding and using them. The system operates in two modes: “playing” and “working”. In the playing mode, graph drawing heuristics (coded as agents in an A-team) are applied at random in order to evolve a set of drawings of a given graph. A history describing what heuristic was used in sequence is kept for every drawing. When the A-team stops, the history for the best generated drawing is the pedigree that we are looking for. Next, in the working mode, the system produces drawings for new graphs by applying only the sequence of heuristics described by the pedigree.

The approach was tested using both undirected and directed graphs. The playing mode produces a high quality drawing, but is very timing consuming because it usually creates many intermediate solutions that have bad quality and are thus discarded. So its main advantage is to produce a pedigree. The working mode, on the other hand, does not create a perfect drawing, but is able to generate a reasonably good solution relatively quickly. The authors suggest that a pedigree could be applied as a recipe to draw similar graphs, but this was not fully tested.

At last, Niggemann and Stain [19] presented an approach for learning the best method for drawing clustered graphs. With this aim in mind, the authors used a set of graphs and a set of popular graph drawing methods. They applied clustering

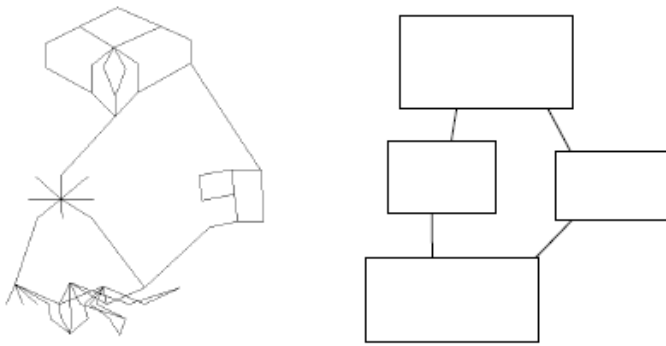


Figure 9. A graph, on the left, and the abstract view its clusters, on the right (from [19]).

algorithms to the graphs and, for each resultant subgraph (cluster), all drawing methods in the set were tested. Next, a knowledge base was created containing the characteristics extracted from all the subgraphs created and the method that generated the best drawing for them. After the learning step, the knowledge base could be used to draw any new graph. This involves partitioning the new graph into clusters, extracting their characteristics and generating a drawing for every cluster using the best method indicated by the knowledge base. The final step is to integrate the drawing in order to create a drawing of the whole graph. Figure 9 illustrates an input graph (the left image) and how it was clustered into four subgraphs (represented by rectangles on the right). Each subgraph has its own characteristics and is drawn by a particular method according to the knowledge base.

IV. CHARACTERISTICS OF THE APPROACHES

Table I summarizes eight aspects of the approaches described in the previous sections:

- *Year of publication.*
- *Class of graphs* – the type of graph that can be drawn by the approach. We use the letter “D” to represent directed graphs and “U” for undirected graphs.
- *Goal of the approach* – the existing approaches have two general goals: to *refine the graph drawing problem* (that is, to learn the user’s preferences of drawing aesthetics and thus to provide a drawing that the user may like); and to *improve the convergence* of a graph drawing algorithm towards a high quality solution (in general, by helping to escape from local minima, focusing the algorithm on more important subproblems, tuning the algorithm’s parameters and providing shortcuts).
- *Knowledge source* – this indicates the source from which the approach obtains information for the learning process. We consider here the main classification followed throughout the paper, regarding whether or not to use information coming from *human interaction*. However, we detail the second case by giving two options that better describe the non-interactive types of source: the *graph structure* alone or data collected from the usage of other *graph drawing methods*.
- *Type of interaction* – when the knowledge source is “human interaction”, then this column specifies the

form of interaction that can be performed by the user to provide the system with information. The types of human interaction that we found in the reviewed literature are: evaluating graph drawings (assigning scores or ranking the drawings), selecting some graph drawings from a list, performing manual adjustments (weights of aesthetic criteria or the drawing itself), and providing graph drawing examples with good or bad aesthetics.

- *Learning method* – this column indicates the learning method used in the approach. The following methods appeared in the reviewed literature: neural networks or related algorithms, case-based knowledge representation, and the adjustment of a graph-drawing evaluation function (that measures the quality of a graph drawing) used as a fitness function in an evolutionary graph drawing method.
- *Reusability* - this means that the approach acquires general knowledge that can be applied to the drawing of other graphs.

TABLE I. CHARACTERISTICS OF THE LEARNING APPROACHES.

Paper	Year of publication	Class of graphs	Goal	Knowledge source	Type of interaction	Learning method	Reusability
			Refining the problem Improving convergence	Data structure Drawing method Human interaction	Evaluation Selection Manual adjustment Drawing examples	Neural network Case-based know. Fitness evolution	
[5]	1994	U/D	X	X	X X	X	X
[7]	1994	D	X	X	X	X	X
[13]	1996	U	X	X		X	
[17]	1998	U	X	X		X	
[8]	1999	U	X X	X	X		X
[18]	1999	U/D	X	X		X	X
[19]	2000	U/D	X	X		X	X
[9]	2001	U	X	X	X		X
[16]	2005	U	X	X		X	
[10]	2013	U	X X	X	X		X
[11]	2014	U/D	X	X	X X		X

From the table, we can see that machine learning techniques have not yet been fully explored to solve problems in graph drawing. There is a major concentration of approaches that use human-computer interaction to adjust a fitness function in an evolutionary graph drawing algorithm. However, reusability of the knowledge learned by the approach is still a weakness. Even when the knowledge is reusable, there is no guarantee that the results will be satisfactory. For example, the learning obtained in [18] can be reused for a new graph drawing task, but the quality of the drawings may not be as good as those obtained by drawing the graph via a complete process (in playing mode). Another aspect of the literature review is that we could not find publications about this theme between the years 2006 and 2012.

V. CONCLUSION AND FUTURE WORK

As far as we know, this is the first literature review to identify the state-of-the-art in the use of machine learning techniques to problems in graph drawing. As we could see from the review, only a few studies investigated such topic.

Considering that the graph drawing research area is well established, with an annual international symposium since 1994, this suggests that there is still much to investigate when the aim is to learn knowledge about how to draw a graph computationally.

The characteristics of the existing approaches presented in Table I can be useful for identifying possible research topics. For instance, case-based knowledge representation has been employed only marginally. There is also a lack of an effective method based on a general neural network for learning about how to draw graphs. Furthermore, the reusability of some described methods should be more deeply experimented with and evaluated. Finally, there are some questions on the current topic that still pose significant challenges. For example, “What other types of knowledge can be collected from a user in order to improve a graph drawing process?” and “How can the knowledge about drawing a particular graph be generalized and applied to other graphs?”.

At the moment, we are investigating actions performed by humans when using interactive graph drawing systems, so that the perceptions and the decisions made by these users can be registered and eventually interpreted, modeled and transformed into algorithms.

ACKNOWLEDGMENT

The authors would like to thank the Brazilian state agency FAPEG for supporting this work with a master scholarship and other financial resources that allowed the necessary infrastructure for research. We also thank Professor Leslie Richard Foulds, National Senior Visiting Professor at INF-UFMG, for revising this paper and suggesting important improvements.

REFERENCES

- [1] G. D. Battista, P. Eades, R. Tamassia, and I. G. Tollis, “Algorithms for drawing graphs: an annotated bibliography,” *Computational Geometry*, vol. 4, no. 5, 1994, pp. 235–282.
- [2] “Graphviz - graph visualization software,” URL: <http://www.graphviz.org/> 2015.01.02 [accessed: 2015-01-02].
- [3] “Gephi - the open graph viz platform,” URL: <http://gephi.github.io/> [accessed: 2015-01-02].
- [4] H. A. D. do Nascimento, “User hints for optimization processes,” Ph.D. dissertation, University of Sydney, Australia, November 2003.
- [5] C. F. X. M. Neto, “A layout system for information system diagrams,” Ph.D. dissertation, University of Queensland, Australia, March 1994.
- [6] C. F. X. M. Neto and P. Eades, “Learning aesthetics for visualization,” in *Anais do XX Seminário Integrado de Software e Hardware*, 1993, pp. 76–88.
- [7] T. Masui, “Evolutionary learning of graph layout constraints from examples,” in *Proceedings of the 7th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '94. New York, NY, USA: ACM, 1994, pp. 103–108.
- [8] A. Rosete-Suarez, M. Sebag, and A. Ochoa-Rodriguez, “A study of evolutionary graph drawing,” *Laboratoire de Recherche en Informatique (LRI), Université Paris-Sud XI*, Tech. Rep. 1228, 1999.
- [9] H. J. C. Barbosa and A. M. S. Barreto, “An interactive genetic algorithm with co-evolution of weights for multiobjective problems,” in *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001)*, 2001, pp. 203–210.
- [10] B. Bach, A. Spritzer, E. Lutten, and J.-D. Fekete, “Interactive random graph generation with evolutionary algorithms,” in *Graph Drawing*, ser. Lecture Notes in Computer Science, W. Didimo and M. Patrignani, Eds. Springer Berlin Heidelberg, 2013, vol. 7704, pp. 541–552.
- [11] M. Spönemann, “Evolutionary meta layout of graphs,” *Institut für Informatik, Christian-Albrechts-Universität zu Kiel, Deutsche Nationalbibliothek*, Tech. Rep., 2014, in English, 21 pages.
- [12] “Aviz: Visual analytics project - graphcuisine,” URL: <http://www.aviz.fr/Research/Graphcuisine> [accessed: 2015-01-05].
- [13] R. Cimikowski and P. Shope, “A neural-network algorithm for a graph layout problem,” *IEEE Transactions on Neural Networks*, vol. 7, no. 2, 1996, pp. 341–345.
- [14] Y. Takefuji and K.-C. Lee, “A near-optimum parallel planarization algorithm,” *Science*, vol. 245, no. 4923, 1989, pp. 1221–1223.
- [15] J. J. Hopfield and D. W. Tank, ““neural” computation of decisions in optimization problems,” *Biological Cybernetics*, vol. 52, no. 3, 1985, pp. 141–152.
- [16] R.-L. Wang and K. Okazaki, “Artificial neural network for minimum crossing number problem,” in *Proceedings of 2005 International Conference on Machine Learning and Cybernetics*, vol. 7, 2005, pp. 4201–4204.
- [17] B. Meyer, “Self-organizing graphs - a neural network perspective of graph layout,” in *Graph Drawing*, ser. Lecture Notes in Computer Science, S. Whitesides, Ed. Springer Berlin Heidelberg, 1998, vol. 1547, pp. 246–262.
- [18] J. Stolfi, H. A. D. do Nascimento, and C. F. X. M. Neto, “Heuristics and pedigrees for drawing directed graphs,” *Journal of the Brazilian Computer Society*, vol. 6, 07 1999, pp. 38 – 49.
- [19] O. Niggemann and Benno, “A meta heuristic for graph drawing: learning the optimal graph-drawing method for clustered graphs,” in *Proceedings of the working Conference on Advanced Visual Interfaces*, ser. AVI '00. New York, NY, USA: ACM, 2000, pp. 286–289.

Game Refinement Theory and Multiplayer Games: Case Study Using UNO

Alfian Ramadhan and Nur Ulfa Maulidevi
 School of Electrical Engineering and Informatics
 Bandung Institute of Technology
 Bandung, Indonesia
 email: masphei@gmail.com, ulfa@stei.itb.ac.id

Hiroyuki Iida
 School of Information Science
 Japan Advanced Institute of Science and Technology
 Nomi, Japan
 Email: iida@jaist.ac.jp

Abstract—Game refinement theory has started to provide some interesting tools to measure sophistication of board games, sport games, and video games. In this paper, we apply game refinement theory to UNO[®] card game, from which we identify valuable aspects regarding multiplayer and incomplete information game. Specifically, we analyze game refinement value zone of UNO and reveal recommended number of players to play. Furthermore, we compare the measure of enjoyment between the players. Experiments have been conducted by developing various computer player types and simulating about 1.4 million UNO games. Results show that critical states of the game and number of card played are the important factors and confirm that UNO is best to play with 4, 5, or 6 players. Furthermore, another result shows that the second last and the last player get the most enjoyment out of the game.

Keywords—UNO card game; game refinement theory; multi-player game; incomplete information game.

I. INTRODUCTION

Game refinement theory has been proposed earlier by Iida *et al.* [1] to determine level of sophistication of games. Some applications have already been done, such as in domain of board games [1], for Mah Jong [2], and sports games [3]. Although there are still many types of games to cover, this theory has already performed well, and generalized fundamental concept. By using sophistication measurement, many facts have been revealed regarding changes of attractiveness of games in decades. In fact, there are still some challenging research questions, especially in applying game refinement theory to multi-player and incomplete information game.

Multi-player game is one of important research themes in game domains. Many works in multi-player game regarding incomplete information aspects have been published, such as multi-player algorithms and approaches [4], comparison of algorithms [5], multi-player Go [6], decision algorithms [7], computing equilibria [8], and lower bounds [9]. Moreover, every kind of games is changing historically by years or decades, even multi-player game. For instance, game refinement theory in multi-person and incomplete information of Mah Jong has been proposed [2]. In fact, recent studies in game refinement theory still focused on several types of games. Hence, multi-player and incomplete information game research in broader types of games are still considered as challenging topics to explore using game refinement theory.

In this paper, we extend game refinement theory with the case study of UNO (UNO[®], is a registered trademark of Mattel Corporation) which has been widely recognized as a popular card games. UNO is commonly known as fascinating games and many variants have been developed in many countries. By analyzing game refinement theory, we discover refinement value and sophistication value zone in UNO that

are appropriated, as has been found for other refined games such as chess [1], Mah Jong [2], and soccer [3]. Contribution of this paper indicates a promising concept of game refinement theory to be applied in any games generally.

Basically, there are some interesting aspects of UNO, because it is categorized as multi-player game, regarding impact or feelings of each player during the game. Exploring recommended number of players to play UNO is challenging. Basically, there is a promising idea, proposed in sports games [3], of using game progress to measure difference of impact for each of player. Furthermore, determining players who enjoy the game the most seems essential to us. Later, we propose some measurement, called enjoyment measurement, to analyze the impact of the game on each player. Thus, we pack our main works on this different problems which are exploring refinement value and sophistication measurement zone in UNO, investigating recommended number of players to play, analyzing enjoyment measure which leads to find who are the players enjoying the most the game.

This paper is organized as follows. Section II introduces game refinement theory. Then, Section III discusses UNO card game, its various versions, our UNO program, and our analysis in applying game refinement theory. In Section IV and Section V, we present our experiments and discussions of our explorations and discoveries. Finally, Section VI concludes and describes some future works.

II. GAME-REFINEMENT THEORY

In this section, we first show a basic idea of game-refinement theory, which has been cultivated in the domain of board games. Then, we present the idea to bridge the gap between the board games and sports games based on a model of game progress and game information progress. Moreover, we consider the game progress model of UNO game.

A. Basic idea of game-refinement theory

We give a short sketch of the basic idea of game refinement theory [3]. This section describes the idea based on Sutiono *et al.* [3] and additional knowledge from authors. The “game progress” is twofold. One is game speed or scoring rate, while another one is game information progress with focus on the game outcome. In sports games such as soccer and basketball, the scoring rate is calculated by two factors: (1) goal, i.e., total score and (2) time or steps to achieve the goal. Thus, the game speed is given by average number of successful shoots divided by average number of shoot attempts. For other score-limited sports games, such as Volleyball and Tennis in which the goal (i.e., score to win) is set in advance, the average number of total points per game may correspond to the steps to achieve the goal [10].

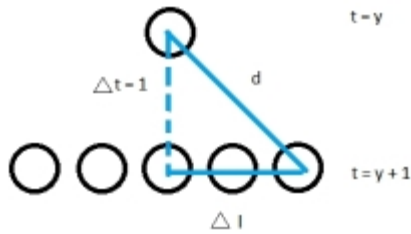


Figure 1. Illustration of one level of game tree.

Game information progress presents how certain is the result of the game in a certain time or steps. Let G and T be the average number of successful shoots and the average number of shoots per game, respectively. If one knows the game information progress, for example after the game, the game progress $x(t)$ will be given as a linear function of time t with $0 \leq t \leq T$ and $0 \leq x(t) \leq G$, as shown in (1).

$$x(t) = \frac{G}{T} t \quad (1)$$

However, the game information progress given by (1) is usually unknown during the in-game period. Hence, the game information progress is reasonably assumed to be exponential. This is because the game outcome is uncertain until the very end of game in many games. Hence, a realistic model of game information progress is given by (2).

$$x(t) = G \left(\frac{t}{T} \right)^n \quad (2)$$

Here, Sutiono et al. [3] described that n stands for a constant parameter which is given based on the perspective of an observer in the game considered. Then, acceleration of game information progress is obtained by deriving (2) twice. Solving it at $t = T$, the equation becomes

$$x''(T) = \frac{Gn(n-1)}{T^n} t^{n-2} = \frac{G}{T^2} n(n-1)$$

It is assumed in the current model that the game information progress in any type of games is happening in our minds. We do not know yet about the physics in our minds, but it is likely that the acceleration of information progress is related to the force in mind. Hence, it is reasonably expected that the larger the value $\frac{G}{T^2}$ is, the more the game becomes exciting due to the uncertainty of game outcome. Thus, we use its root square, $\frac{\sqrt{G}}{T}$, as a game refinement measure for the game considered.

B. Board games and sports games

Here, we show the idea to bridge the gap between board games and sports games by deriving a formula to calculate the game information progress of board games [3]. Let B and D be average branching factor (number of possible options) and game length (depth of whole game tree), respectively. One round in board games can be illustrated as decision tree. At each depth of the game tree, one will choose a move and the game will progress. One level of game tree is illustrated in Fig. 1. The distance d , which has been shown in Fig. 1, can be found by using simple Pythagoras theorem, thus resulting in $d = \sqrt{\Delta t^2 + 1}$.

Assuming that the approximate value of horizontal difference between nodes is $\frac{B}{2}$, then we can make a substitution and get $d = \sqrt{\left(\frac{B}{2}\right)^2 + 1}$. The game progress for one game is the total level of game tree times d . For the meantime, we do not consider Δt^2 because the value ($\Delta t^2 = 1$) is assumed to be much smaller compared to B . The game length will be normalized by the average game length D , then the game progress $x(t)$ is given by $x(t) = \frac{t}{D} \cdot d = \frac{t}{D} \sqrt{\left(\frac{B}{2}\right)^2 + 1} = \frac{Bt}{2D}$. Then, in general we have, $x(t) = c \frac{B}{D} t$, where c is a different constant which depends on the game considered. However, we manage to explain how to obtain the game information progress value itself. The game progress in the domain of board games forms a linear graph with the maximum value $x(t)$ of B . Assuming $c = 1$, then we have a realistic game progress model for board games, which is given by

$$x(t) = B \left(\frac{t}{D} \right)^n. \quad (3)$$

Equation (3) shows that the game progress in board games corresponds to that of sports games as shown in (2).

To support the effectiveness of proposed game refinement measures, some data of games such as Chess and Go [1] from board games and two sports games [3] are compared. We show, in Table I, a comparison of game refinement measures for various type of games. From Table I, we see that sophisticated games have a common factor (i.e., same degree of acceleration value) to feel engagement or excitement regardless of different type of games.

TABLE I. MEASURES OF GAME REFINEMENT FOR BOARD GAMES AND SPORTS GAMES.

Game	B or G	D or T	$\frac{\sqrt{B}}{D}$ or $\frac{\sqrt{G}}{T}$
Chess	35	80	0.074
Go	250	208	0.076
Basketball	36.38	82.01	0.073
Soccer	2.64	22	0.073

III. UNO CARD GAME

UNO is one of the well known card games in the world and characterized as a multi-player, imperfect-information, and uncooperative combinatorial game [11]. In addition, a poll found on the website BoardGameGeek, a website specialised on board games and card games, shows that UNO is recommended to play with 2 to 10 players and best to play with 4, 5, or 6 players [12].

Research of UNO card game has attracted many people globally. Recent works have been performed from the viewpoint of a combinatorial algorithmic game theory [11], also in playful probing [13], and intelligent system [14]. Thus, UNO has been recognized not only in entertainment, but also in academic domain.

There are many variants of UNO with different rules, number of cards, or number of players which can be found in various countries in the world. Pagat [15] is a website which collects information of UNO variants including Hold'em UNO, Magic UNO, Speed UNO, Solitaire UNO, and so forth.

There is also a modified version of UNO rules which is played slightly different. The only one difference is that the game ends until there is only one player left who still holds cards in hand. This type of game is similar with DaiFuGo [16]

card game from Japan. Moreover, the modified rules of UNO are used in our experiments to measure enjoyment value of each player.

A. Basic rules

UNO official rules can be found at official site [17] or [18]. There are 108 cards in total which are organized as follows: 19 Blue cards, 19 Green cards, 19 Red cards, 19 Yellow cards, 8 Draw Two, 8 Reverse, 8 Skip, 4 Wild, and 4 Wild Draw Four. Accordingly, Draw Two, Reverse, Skip, Wild, and Wild Draw Four cards are defined as Action cards which have effects as they are played in the game. Object of the game is to be the first player in games to get score 500 points. Specifically, only the winner gets score from a game by getting rid of all the cards in hands before other players, and this score is calculated from all of opponents' cards left.

Basically, UNO is easy to play. First of all, the game begins by deciding who among participants is to play first. In this part, every player picks a card and the first player is determined by the one who gets the highest number of numbered cards. Then, each player when beginning his turn firstly has to determine whether he wants to draw a card, or play a card in his hand. He can choose to play a card in his hand, otherwise he draws a card from deck and can play the drawn card if the card is possible to play. In official rules, the game ends until there is one player which has no cards left in his hands. However, we add some modifications in this study that the game ends until there is only one player holding cards in his hands. By doing so, there will be ranking from first player as the winner until the last player as the final loser.

There are several action cards which have to be understood before playing such as Draw Two, Skip, Reverse, Wild, and Wild Draw Four. Draw Two card forces the next player to draw two cards and skip his turn. Skip card means that next player misses his turn. Reverse card is used to invert turn direction. Wild card can be used to change color to play. Then, Wild Draw Four card is used to force next player to draw four cards, skip next player's turn, and change color in the game. These cards have their own effect and affect game play. Thus, these basic rules and action cards lead analysis of UNO in next sections.

B. Game refinement theory and game progress in UNO

The idea that had been the basis of previous works on sports games [3] is to find some critical enjoyment points in the game, and only measure those, assuming that they are key point and are the only point that we need to study. For example, in soccer, this critical action are the shoots. A game of soccer is actually more than a succession of shoot, but we can restrict our study to only those for two reasons: the first is all the other actions during the game take place only to decide which side will be able to try a shoot, and how probably it will be a success, and the second is that shoots are the moment the spectator can enjoy the most, because it is the most intense action.

In game refinement theory, branching factors and game length are the main factors to determine game information outcome [1]. Iida *et al.* proposed average number of possibilities and turns to apply game-refinement theory in board games [2]. Furthermore, Sutiono *et al.* proposed some relationships between game-refinement theory and game progress concluding that number of goals and shoots are factors to measure sophistication of games, as well as game information

outcome [3]. Thus, each game may have its own measurement to be identified as game refinement value regarding game characteristics.

UNO card game is different from sports and board games. Although there are some similarities between board games and card games such as turns and type of actions, different rules or characteristics of games can result contrast impact to players in terms of game refinement theory. For instance, different versions of Mah Jong game in history affect its attractiveness [2]. Furthermore, broad and deep analysis of UNO are required since it is characterized as a multi-player, imperfect-information, and uncooperative combinatorial game [11]. Thus, there are some different considerations to identify main factors of game refinement theory in UNO.

In this study, we highlight multi-player and imperfect-information characteristics as main aspects. First of all, multi-player games characteristic is simply identified by the number of players. Basically, each player is supposed to perform any actions which affect the game in any conditions. For instance, a player may play any action cards to attack other players or skip his turn to give other player's chance. Consequently, each player has contributions to increase or decrease attractiveness or flow of the game. Meanwhile, treating imperfect-information game is more challenging than perfect-information game because of hidden information. However, there is global information which is visible to all of players in game and simply measured as global variables which is state of the game such as number of cards or number of remained cards in deck.

Because of the imperfect information nature, it is much harder to measure the progression of the game. The end of the game may not be expected by every player. We will only look at information that are shared by every player for our measurement, and because the game ends when a player had no more cards in his hand, game progress can be measured by looking at the number of remaining cards in hand for each player.

The solution we propose here is to consider the times a player say "UNO", when he has only one card left in hand, which we name critical state of the game. This is a viable option because it is emphasised by the game itself as a big point of interest of the progression of a player toward victory. When a player says "UNO", he is more likely to become the target of every punishing card or strategy from other players (for example playing yellow if it is known that he had no yellow card in hand). Also, when considering the ratio of "UNO" over the total number of card played, the value obtained is in relation with the balance of the game. The most card you need to play in average before having only one card left is a measure of how slow the game is.

In experiment, we use both values in average of each player because UNO is a multi-player game. Thus, determining sophistication of the game is simplified by using average number of UNO times and average number of played cards for each per player.

$$x(t) = U \left(\frac{t}{P} \right)^n \quad (4)$$

According to (3), let U to replace B as average number of saying UNO, and P to replace D as average number of played cards for each player independently. By using our analysis and referring to game refinement theory, game information outcome of UNO is defined in (4). Then, acceleration of game

information outcome value in UNO is shown in (5). Finally, the sophistication measurement can be obtained using $\frac{\sqrt{U}}{P}$.

$$x''(t) = \frac{Un(n-1)}{P^n} t^{n-2} \quad (5)$$

When $t = P$, the equation becomes

$$x''(P) = \frac{U}{P^2} n(n-1)$$

On the other hand, there is another analysis regarding excitement for each player which argues that the first player, the winner, does not feel more excitement than other late player. Likewise, there is some different feeling when comparing multi-player game with two-player game due to the number of players. Although the first player is lured by prize and high score, as he enjoys his win, but it stops his play. But players still in the game can continue to enjoy it. Our model does not include feeling about winning, and only focus on enjoying the content of the game, not the side effects like winning or losing. Our analysis argues that the second and the last player in the game are the player who enjoy the most and feel the most excitement in the game.

Basically, measuring enjoyment feeling is more likely to outlook the game in overall. This can be done by using overall number of critical states of the game and contributions of the game as well as generalizing our analysis in measuring sophistication value of UNO. In other words, overall critical state of the game is reflected by total number of UNO times in the game. Moreover, instead of having overall number of played cards, we can change it by number of rounds in the game. Thus, according to the sophistication measurement formula, the enjoyment measure can be similarly defined using $\frac{\sqrt{U}}{P}$ with U as overall UNO times and P as number of rounds, in order to specifically investigate the excitement of each player.

C. UNO program

We have created a program which is developed in Java to run our simulation of UNO. The program works as automatic simulation playing UNO and records each player's activity. Information of player's activities is collected during the game including the number of turns, played cards, UNO times, and so forth. Furthermore, the program has been published as an open source which can be found on Github [19].

Process of each player who is having a turn is illustrated in Fig. 2. Square and diamond shape stands for process and decision, respectively. From the beginning until the end of turn, a player is given some several processes and decisions including decision of drawing a card, process of picking a card, and so forth. In general, there are several basic actions called Module Actions which require an action of each player regarding their strategy and mind. These actions are drawing a card, playing drawn card, playing a card, yelling UNO, and choosing a color. Consequently, by having well defined and separated actions, the program becomes well structured and modularized, especially in building computer players.

Basically, our simulation is played fully by computer players which are inspired from multi-player algorithms and approaches [4]. However, our implementation has been performed in simpler ways. We have created four different profiles as computer players: Amateur, Offensive, Defensive, and

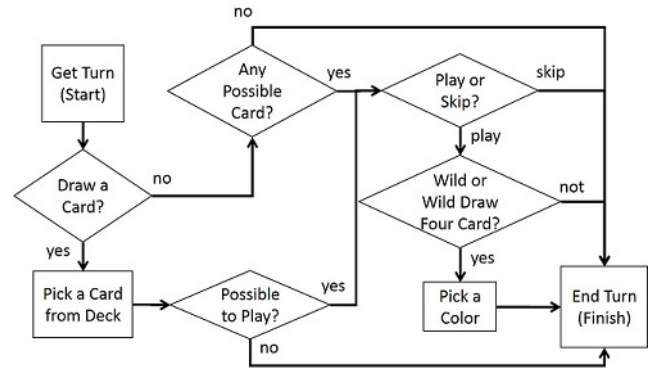


Figure 2. Flowchart of Player Turn Process.

Observer. These types of players are identified because they are most reasonable strategies and easy to understand.

First of all, Amateur is the easiest player to imagine. Amateur is more likely to be analogous to weak human players who still lack of experiences. In general, Amateur does not have good strategy to play, so that he does not consider any actions or situations. Specifically, Amateur plays all of Module Actions randomly and recklessly. Thus, action of Amateur is not specified as an important algorithm.

Algorithm III.1: OFFENSIVE(*possible_cards*)

```

procedure OFFENSIVE_ACTION(P)
  C ← {} // candidate stack
  C ← TOP(P) // P possible cards as stack
  for each p ∈ P
    if (stronger (p, TOP(C)))
      do {
        // p is stronger
        // push p to top of stack C
        PUSH(C p)
      }
  return (C) // return cards in offensive order
  
```

Figure 3. Offensive step.

Secondly, Offensive roles as an active player and is more likely to be analogous to ambitious players. In general, Offensive always plays offensively, so that he takes the most strongest and offensive actions, especially to attack other players as shown in Fig. 3. Ranking of card's strength from the strongest to the weakest is given as follows, Draw Four, Draw Two, Skip, Reverse, and Wild card.

Meanwhile, Defensive player acts as a passive player. In contrast, Defensive is more likely to be the opposite of Offensive because he always plays defensively, so that he mainly chooses harmless actions as shown in Fig. 4. Ranking of cards is prioritized in inverted order with Offensive algorithm.

Finally, Observer player chooses his strategy by considering opponent actions. Basically, he plays cards which other players do not have the color or number by remembering others' missing cards from recent turns, especially when they draw a card as shown in Fig. 5.

Algorithm III.2: DEFENSIVE(*possible_cards*)

```

procedure DEFENSIVE_ACTION(P)
  C ← {} // candidate stack
  C ← TOP(P) // P possible cards as stack
  for each p ∈  $\mathcal{P}$ 
    if (weaker (p, TOP(C)))
      do {
        if (weaker (p, TOP(C)))
          then {
            // p is weaker
            // push p to top of stack C
            PUSH(C, p)
          }
      }
  return (C) // return cards in defensive order

```

Figure 4. Defensive step.

Algorithm III.3: OBSERVER(*possible*, *forbidden*)

```

procedure OBSERVER_ACTION(P, F)
  C ← [] // list of candidates

  // copy P into C
  for each p ∈  $\mathcal{P}$ 
    do {INSERT (C, p)}

  for each f ∈  $\mathcal{F}$ 
    do {
      for each p ∈  $\mathcal{P}$ 
        do {
          if match (f, p)
            then {VOTE_DOWN(C, p)}
        }
    }

  SORT(C) // sort candidates in vote order
  return (C) // return cards in vote order

```

Figure 5. Observer step.

IV. EXPERIMENTAL DESIGN AND RESULT

In this paper, we conduct experiments by simulating our UNO computer program to obtain refinement value and sophistication measurement. Then, we quickly compare a particular result with the real UNO games played with human players. Another experiment is done by modifying rule of the game to identify enjoyment value using UNO computer program.

A. Game Refinement Experiment

First of all, we collected data from 1,432,089 game simulations run by several type of players described previously. Composition of player types in each of the game is randomly organized. Measures of game refinement in simulations of UNO is illustrated in Table II. The measures are applied for 10 different number of players playing in the game from 3 to 12 players. Each number of players gives a different value of three variables which are average UNO times per player U , average played cards per player P , and division between square root of U and P as game information outcome.

According to Table II, U , P , and $\frac{\sqrt{U}}{P}$ are decreasing from 3 to 12 players. That is, chance to have UNO is decreasing when the number of players is increasing. In addition, each player also has less number of played cards in the game with more players. Furthermore, third variable called sophistication

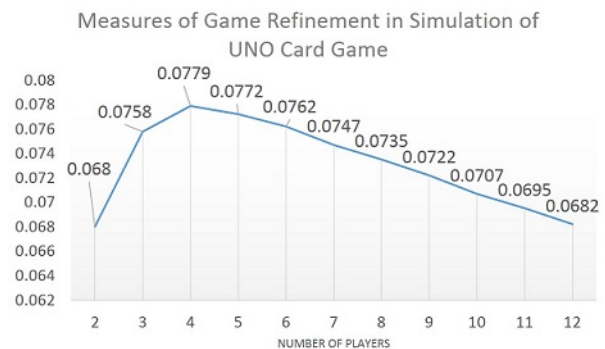


Figure 6. Graph of Game Refinement Measurement UNO Simulation.

measurement is scored 0.0682 as minimum value and 0.0758 as maximum value. Specifically, sophistication measurement reaches below 0.07 in 11 and 12 players. Understandable visualization can be seen from Fig. 6, which shows visualization of Table II.

TABLE II. MEASURES OF GAME REFINEMENT IN SIMULATION OF UNO CARD GAME.

Player	U	P	$\frac{\sqrt{U}}{P}$
3	1.260	14.802	0.0758
4	0.976	12.684	0.0779
5	0.813	11.679	0.0772
6	0.702	10.994	0.0762
7	0.617	10.511	0.0747
8	0.554	10.129	0.0735
9	0.506	9.856	0.0722
10	0.460	9.594	0.0707
11	0.427	9.404	0.0695
12	0.398	9.246	0.0682

On the other hand, although our experiments using simulation show fascinating results, statistical data which was obtained from real games shows slight difference. By conducting 19 real UNO games in four human players, we have U at 1.118 and P at 10.947, so that the $\frac{\sqrt{U}}{P}$ becomes 0.095. This phenomenon is interesting to be explained in discussion section later on.

B. Enjoyment Value Experiment

We conduct another experiment by using UNO modified rules, so that the game continues until only one player had cards in hand. The results shown in Table III and Table IV illustrate measures of game refinement in simulation of UNO with modified rules. The measurement is performed for 10 different number of players from 3 to 12 players. Ranking are grouped by number of players in the game, and the enjoyment value is expressed in function of the rank of the player.

According to Table III and Table IV, score of each player is decreasing from first rank until last rank in the game, so that the first rank gets the highest score and the last rank gets the second lowest score. Specifically, the lowest and highest score in each different player of the game is from about 0.074 to 0.076 and 0.11 to 0.22, respectively. Basically, only the second and last player from all of different number of players perform score from about 0.071 to 0.076.

There are some essential information extracted from statistics regarding different type of computer players. Other statistics from our simulation of UNO with modified rules in overall

TABLE III. ENJOYMENT MEASURE OF UNO SIMULATION WITH MODIFIED RULES 3-7 PLAYERS.

Rank	3P	4P	5P	6P	7P
1	0.110	0.135	0.152	0.166	0.177
2	0.071	0.106	0.130	0.148	0.164
3	0.074	0.072	0.103	0.126	0.144
4		0.074	0.072	0.102	0.123
5			0.074	0.073	0.101
6				0.075	0.074
7					0.075

TABLE IV. ENJOYMENT MEASURE OF UNO SIMULATION WITH MODIFIED RULES 8-12 PLAYERS.

Rank	8P	9P	10P	11P	12P
1	0.188	0.197	0.205	0.213	0.220
2	0.178	0.191	0.202	0.212	0.222
3	0.160	0.174	0.187	0.199	0.210
4	0.141	0.156	0.170	0.183	0.195
5	0.121	0.138	0.153	0.167	0.180
6	0.100	0.119	0.135	0.150	0.164
7	0.074	0.098	0.117	0.134	0.148
8	0.075	0.074	0.098	0.116	0.132
9		0.075	0.074	0.097	0.115
10			0.076	0.075	0.097
11				0.076	0.075
12					0.076

is collected in Table V. Statistics are given as a percentage and score from all of different type of players, which are Amateur, Defensive, Offensive and Observer. Furthermore, according to Table V, the lowest and highest percentage of winning rate are performed by Amateur at 10.04% and Defensive at 30.5%. In addition, Amateur and Defensive also perform the lowest score at 0.0862 and the highest score at 0.1337 of enjoyment measure, respectively. On the other hand, Amateur reaches the highest percentage of being second last player at 38.42% and being last player at 57.82%.

TABLE V. STATISTICS OF UNO.

Level	Amateur	Defensive	Offensive	Observer
Winning rate	10.04%	30.50%	29.62%	29.84%
Being second last player	38.42%	21.55%	21.21%	18.82%
Being last player	57.82%	13.35%	13.84%	14.99%
enjoyment measure	0.0862	0.1337	0.1323	0.1246

Enjoyment measure of 8-Player UNO is represented in Fig. 7 ordered by ranking. First rank player gets the highest score at 0.188, but the second last player gets the lowest score at 0.074. In general, the point is decreasing from the first rank until the second last player. Meanwhile, the second and the last player perform score 0.074 and 0.075, respectively.

V. DISCUSSION

In this section, there are discussions regarding refinement value of UNO and enjoyment measurement in the game according to the experimental results. First investigation focuses on the difference of refinement value between UNO computer program and the real game with human players. Then, the second issue points the enjoyment value of the game which is specifically related with the second and the last player in the game.

First of all, by observing comparison of game refinement measurement between computer simulation and human UNO games, there is difference about 0.02. The difference is shown by 4-Player human game and 4-Player UNO game simulation

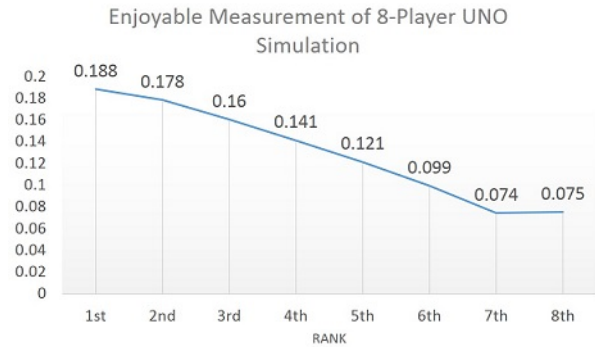


Figure 7. Enjoyment Measure of 8-Player UNO.

in Fig. 6. There are several possible issues to be drawn regarding this difference such as computer player quality, playing against human in reality, and method of gathering data.

Basically, our implementation of UNO players, which are Amateur, Offensive, Defensive, and Observer may not fully represent real human players' abilities. They are only very simple models, and could be improved to be closer to human real strategies. Real human players' abilities may vary broader and deeper in terms of skill compared to our implementation. Moreover, playing UNO against human players is more likely to be uncontrollable, so that the game can be various depending on various human skills. For example, human social nature may cause players to not be equally considered by each other. Besides that, another problem may be drawn by error in recording data in human game. For instance, gathering data in real human game may be less accurate because of various game flow speed. There are many factors affecting the game flow such as luck, number of turns, strategies and so forth. However, our implementation has reflected only a few of the various player types in the world. Although the difference appears, our implementation is fair enough since the difference is not significant. Furthermore, the number of data obtained from human real games is less than 20, which is very small compared to our computer simulation with only slight difference of result. Thus, the statistics gathered from computer simulation is still acceptable to be analyzed.

By accepting data of our simulation, we can analyze from Fig. 6 that UNO shows refinement value from 0.0682 to 0.0758 for different number of players. In fact, there is an interesting fact that the value reaches lower than 0.07 for 11 or 12 players. By referring to game refinement theory in board games [2] and sports games [3], we can say that UNO is sophisticated enough to play from 3 to 10 players since the refinement value ranges in between 0.07 and 0.08, which are identified as the reasonable values. Moreover, although our experiment did not cover 2-player game, we have successfully confirmed that recommended number to play UNO ranges between 2 and 10 players, which is what people voted on BoardGameGeek. Our measures shows that UNO is a sophisticated game to play, like chess, Mah Jong and soccer. Furthermore, we suggest that the most sophisticated game to play UNO is not more than 10 players.

In addition, according to Fig. 6, there is a peak area which is considered as the three highest refinement value in UNO, which are performed by 4, 5 and 6 players. By referring to

game refinement value, this can be inferred that these three cases are special numbers of players which is recommended to enjoy the most the game of UNO, because it has the best sophistication for these values. This finding is relevant to the BoardGameGeek site, where people have voted optimal numbers to play UNO, as well as the same result in our experiment.

Moreover, according to our last experiment in Fig. 7, there is another important result showing about enjoyment value in the game. Although the last player performs higher score than the second last player, their values are in the range of 0.07 to 0.08. According to related works, by using board and sports game's sophistication value which lies between 0.07 and 0.08, our enjoyment measure shows that late players feel more excitement comparing to players who left the game earlier, especially the first player. Hence, we have confirmed that the second and the last player enjoy the most in the game. Finally, these results endorse ideas to support the concept of game refinement theory to be applied generally.

VI. CONCLUSION AND FUTURE WORK

Game refinement theory has been applied to measure sophistication of games in board games, video games, and sports. In this paper, we have extended this theory to card games and presented that UNO can be analyzed using this measurement. Specifically, we can prove the recommended number of players to play UNO and show some enjoyment measure to determine which player enjoy the most in the game.

In UNO card game, individual critical situation and individual contributions are the most considered values to be used in sophistication measurement in game refinement theory. In addition, the theory considers UNO card game as a sophisticated game which is consistent with the popularity of the game. Furthermore, recommended number of players to play UNO has been proven from 3 to 10 players and the best number to play lies from 4 to 6 players as mentioned on BoardGameGeek. Thus, game refinement theory is well applied in UNO card game using the individual critical situation and contribution.

Determining which player who feels the most engaging game is a challenging question to be figured out, especially in multi-player games. Directly, enjoyment measure can be simplified using critical states of the game and game length in modified UNO card game. Specifically, this variables show some facts that the second last and the last player enjoy playing the most in the game. Consequently, the critical state of the game and length of the game perform the most important role in identifying the enjoyment measure of player in the game.

Thus, a good deal of efforts have been done to analyze UNO card game using game refinement theory approach. This paper has successfully proposed attributes which can be used as sophistication and enjoyment measure in UNO which are individual critical situation, contributions, overall critical states of the game, and game length. In general, critical state of the game is reasonably a main factor in game refinement theory, especially in multi-player games in order to discover sophistication or enjoyment evaluation of games.

This research can be continued better by exploring external validation to discover fundamental formulas in game refinement theory. Besides, it is possible to capture whole picture of game in general by inspecting carefully all of the applied concepts so far and identifying global concept of game. Moreover, further works may consider other interesting aspects

such as cooperation and non-cooperation in multi-player and incomplete information games using game refinement theory. In addition, improving computer player UNO in terms of quantity and quality may also be interesting, especially in developing a framework of more or less sophisticated multi-player game. Moreover, there are other challenging aspects to apply game refinement theory in multi-player and incomplete information games such as player modelling, social behavior, economy, and game sustainability.

REFERENCES

- [1] H. Iida, N. Takeshita, and J. Yoshimura, "A metric for entertainment of boardgames: Its implication for evolution of chess variants," *Entertainment Computing Technologies and Applications*, pp. 65–72, 2003.
- [2] H. Iida, H. Takahara, J. Nagashima, Y. Kajihara, and Y. Hashimoto, "An application of game-refinement theory to mah jong," in *International Conference on Entertainment Computing (ICEC)*, pp. 333–338, 2004.
- [3] A. P. Sutiono, A. Purwarianti, and H. Iida, "A mathematical theory of game refinement," *6th International Conference on Intelligent Technologies for Interactive Entertainment (INTETAIN 2014)*, 2014.
- [4] N. R. Sturtevant, *Multi-Player Games: Algorithms and Approaches*. PhD Thesis, University of California, Los Angeles, 2003.
- [5] N. R. Sturtevant, "A comparison of algorithms for multi-player games," in *Third International Conference on Computers and Games*, pp. 108–122, 2002.
- [6] T. Cazenave, "Multiplayer go," in *LNCS 5131*, pp. 50–59, 2008.
- [7] G. Peterson, J. Reif, and S. Azhar, "Decision algorithms for multiplayer non-cooperative games of incomplete information," *Journal of Computers and Mathematics with Applications*, vol. 43, pp. 179–206, 2002.
- [8] S. Ganzfried and T. Sandholm, "Computing equilibria in multiplayer stochastic games of imperfect information," in *IJCAI'09 Proceedings of the 21st international joint conference on Artificial intelligence*, pp. 140–146, 2009.
- [9] G. Peterson, J. Reif, and S. Azhar, "Lower bounds for multiplayer non-cooperative games of incomplete information," *Journal of Computers and Mathematics with Applications*, vol. 41, pp. 957–992, 2001.
- [10] R. Takeuchi, R. Ramadan, and H. Iida, "Game-refinement theory and its application to volleyball," *IPSI SIG Technical Report*, 2014-GI-31, 3, pp. 1–6, 2014.
- [11] E. D. Demaine, M. L. Demaine, R. Uehara, T. Uno, and Y. Uno, "Uno is hard, even for a single player," in *FUN'10 Proceedings of the 5th international conference on Fun with algorithms*, pp. 133–144, 2010.
- [12] S. Alden and D. Solko, "Boardgamegeek." url: <http://boardgamegeek.com/boardgame/2223/uno>. Accessed: 15-02-2014.
- [13] R. Bernhaupt, A. Weiss, M. Obrist, and M. Tscheligi, "Playful probing: Making probing more fun," in *INTERACT'07 Proceedings of the 11th IFIP TC 13 international conference on Human-computer interaction*, pp. 606–619, 2007.
- [14] L. Antanas, I. Thon, M. van Otterlo, N. Landwehr, and L. De Raedt, "Probabilistic logical sequence learning for video," in *Proceedings of the 19th International Conference on Inductive Logic Programming*, 2007.
- [15] J. McLeod, "Card game rules." url: http://www.pagat.com/invented/uno_vars.html. Accessed: 15-02-2014.
- [16] J. McLeod, "Rules of card games: Dai fugo / dai hinmin." url: <http://www.pagat.com/climbing/daifugo.html>. Accessed: 22-01-2015.
- [17] UNOTips.org, "Uno card game official rules from unotips.org." url: http://unotips.org/pdf/official_rules.pdf. Accessed: 15-02-2014.
- [18] E. D. Demaine, M. L. Demaine, N. J. A. Harvey, R. Uehara, T. Uno, and Y. Uno, "The complexity of uno," *arXiv:1003.2851*, 2010.
- [19] A. Ramadhan, "Uno card games simulation using java." url: <https://github.com/masphei/unosuka>. Accessed: 22-01-2015.

ERP Implementation in a Developing World Context: a Case Study of the Waha Oil Company, Libya

Hosian Akeel
Department of Computing
Azzaytuna University
Libya
Husage_81@hotmail.com

Martin Wynn
School of Computing and Technology
University of Gloucestershire
Cheltenham, UK
MWynn@glos.ac.uk

Abstract – This article examines the implementation and functioning of a major Enterprise Resource Planning software package in a Libyan oil company. The study uses a process mapping and systems profiling approach to establish the current status of a 10 year project - which the company embarked upon in 2007 - to implement the new company-wide software. It examines the project from the point of view of progress in new technology, new people skills and process change, and concludes that a balanced approach to these three change elements has underpinned successful project outcomes.

Keywords – Enterprise Resource Planning; Libyan oil companies; information systems; ERP; process change; information systems strategy.

I. INTRODUCTION

It has been over 25 years since Enterprise Resource Planning (ERP) packages first came to the market. The emergence of the UNIX operating system in the late 1980s as a standard for minicomputers running the Intel chipset significantly increased demand for integrated packaged software. SAP and Oracle were two of the first software vendors to offer packages that combined software modules, providing integrated solutions for order capture and processing, invoicing, ledgers, materials requirements planning, inventory management, payroll, and human resource management. The functionality and integration of this software was enhanced in the 1990s, sometimes through the acquisition and incorporation of rival vendors' software. By the end of the decade, ERP offered the prospect of one integrated package for all company operations. As Koch noted, "ERP attempts to integrate all departments and functions across a company on to a single computer system that can serve all those departments' particular needs" [1].

ERP packages were initially used in some of the major international organizations of the Western world. In the UK, for example, GlaxoSmithKline, Kraft, Nestle, Kellogg's and Diageo all became early users of the SAP ERP package. The number of ERP vendors increased as small to medium sized enterprises started to acquire and deploy cheaper ERP packages, specifically geared to smaller scale operations. In the developing world, the uptake of these new systems was slower, for a number of reasons, including budgetary constraints of the user organization, and the non-availability of sales and support operations for many of these vendors in developing world countries. Since the turn of the new

millennium, the use of ERP in developing world countries has accelerated, but the current literature suggests that there have been both significant failures [2], as well as successes [3]. Although many authors [4] [5] now question the concept of the "digital divide" between developed and developing worlds, there remain divergences of opinion regarding the suitability of systems developed in the western world for a developing world context. When discussing IS in the developing world, Gomez and Pather [6] observe that there is lack of literature and evaluation studies, and the World Bank view that "analysts and decision makers are still struggling to make sense of the mixed experience of information technologies in developing countries" is highlighted by other authors [7]. This article attempts to contribute to addressing this imbalance and focuses on the implementation of SAP – a mainstream ERP package - in the Waha Oil Company (WOC) in Libya, identifying the key issues that underpinned project outcomes.

This introductory section is followed by a discussion of the theoretical framework for this paper in Section 2. In Section 3, the case study methodology used in this research is briefly discussed, and Section 4 then focusses on the primary research findings from the core study of the ERP implementation in WOC. In Section 5, the benefits that have ensued from the company's information systems (IS) strategy are outlined, highlighting the importance of making sure that change in process, people and technology is kept in balance, as large IS projects such as this are progressed.

II. THEORETICAL FRAMEWORK

The Design - Actuality Gap model developed by Heeks [8] identifies four main elements of change that are key to transitioning an organisation from local actuality - where the organisation is now – to its future state or design (see Figure 1). While Heeks' model can be applied to various business change environments, in this paper it is used to support the analysis of the implementation of an ERP package. Other authors [9] have adopted a similar approach in looking at structures that are embedded in both packages and organisations, in trying to assess the reasons for misalignments between an ERP package and the organisation. According to Heeks, the transition to such a major new system can be considered from four interrelated dimensions of change – people, structure, technology and processes. Here, however, we will focus on what have been

termed the “three pillars” in other literature [10] for successful systems implementation – people, processes and technology. This is justified because the structure of the company under study changed little over the period of ERP implementation, although processes within the company were affected. As regards process change, Harmon [11] has argued that process redesign should not only look at the top level process functions, but should also examine how the lower level activities are managed day-to-day, looking at how activities are planned, communicated, organised, monitored, and controlled.

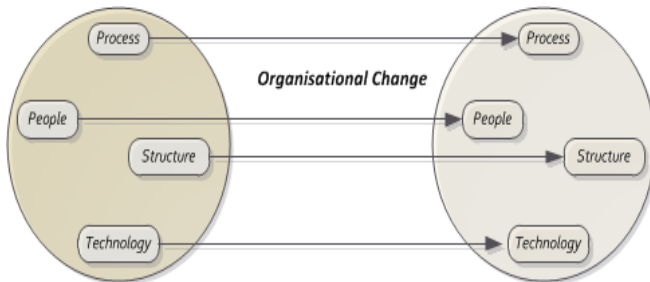


Figure 1. Design - Actuality Gap Model
Source: Adapted from Heeks [8]

The process mapping technique can help the researcher assess systems deployment at process level. It generates a sequence of maps that are used in identifying the information systems that are used in defined business areas. While process mapping is used as a framework to identify the business processes, it can also be used as a point of reference for assessing the functionality of the information systems themselves at process level. This “systems profiling” encompasses a review and assessment of functionality, reporting capabilities, user interface and soundness of the underlying technology [12].

Within this context, and in accordance with the research aims and objectives given above, this research addresses the following questions:

1. How has the SAP ERP package been implemented at WOC and what is the underlying IS strategy?
2. How successful has the implementation been in supporting the growth of the company?
3. How can the ERP implementation at WOC be assessed using the three pillars of people skills enhancement, technology capability improvement, and process change?

III. RESEARCH METHOD

A case study approach is adopted in this research. “A case study is a research strategy that focuses on understanding the dynamics present within single settings” [13]. The case study is a widely used research method within business research, and can focus on one single case or a single organization [14]. Bryman and Bell [15] argue that the case study is particularly appropriate to be used in

combination with a qualitative research approach, allowing detailed and intensive research activity. In a qualitative case study, an inductive approach is usually adopted, and a case study is also appropriate for the combination of qualitative methods. This is particularly relevant to this study of information systems in one large oil company, which combines mapping and profiling techniques with questionnaire and interview material. Saunders, Lewis and Thornhill [16] argue that case studies are mostly used in explanatory and exploratory research, and affirm that the case study is a good choice when the researcher requires a deep understanding of one specific organisation.

Case studies are often used to provide insights into a significant managerial issue, providing an analysis of the context and processes that illuminate the theoretical issues being studied [17]. The need for a case study arises from the quest to understand, analyze, and highlight a complex phenomenon and it is explicitly applicable for doing research on strategic developments or processes [18]. The primary purpose for undertaking a case study is to explore the particularity, the uniqueness, of the single case. The goal for the researcher is to design a good case study and to collect, present and analyze data “fairly” [19].

The central tendency among all types of case study is that it tries to illuminate a decision or set of decisions: why they were taken, how they were implemented and with what result. Because phenomenon and context are not always distinguishable in real-life situations, other technical characteristics, including data collection and data analysis strategies, have to be recognized as a secondary technical dimension of case studies. In this context, case studies are not limited to being a data collection tactic alone or even a design feature alone. Like other research methods, it is a way of investigating an empirical topic, by following a set of specific procedures to gain appropriate valuable information and ideas [18].

In this research, the case study under investigation is the SAP implementation at WOC, the largest domestic oil company in Libya. It was founded in 1955, and its headquarters are in Tripoli, the Libyan capital city. It is 59% publicly owned (by Libyan state entities) and 41% privately owned. It employs 3,200 staff, has over 1000 oil wells and an annual turnover of 690 million Libyan dinars (LDs) in 2013 (1 Libyan dinar = \$0.8US). Since 2007, the company has spent over 1.5 million LDs on hardware and 4 million LDs on software. The main investment in software has been on SAP licence acquisition and annual maintenance charges.

Data collection was pursued through a combination of questionnaires, interviews, observations and documentary evidence. Yin [18] suggests that the utilisation of multiple sources of evidence is one way of increasing the construct validity of case studies. A detailed structured questionnaire was filled in by two respondents – the IT manager and the Finance Director, and this was followed up by face to face interviews to build up a picture of IS strategy and systems deployment in recent years. There were several iterations of follow-up emails and phone calls to clarify points made in the questionnaire responses, and this resulted in further

conversations with other staff members. The topics included in the questionnaire can be categorised as follows:

- a) Company information: To confirm basic company data, company profile, size, turnover, operations and other general information.
- b) Company processes: To explore the company's main business processes and to determine secondary processes (sub-processes).
- c) Information systems: To establish the deployment of information systems and to assess the underpinning technical architecture.
- d) Current systems status: To confirm the functionality of the main information systems and general satisfaction levels in different departments that use them.
- e) Problems and challenges: To determine if there were any key problems or issues, both from a technical perspective and also from the point of view of the end user. Integration and interfacing of systems, report quality, systems performance and access were some of the issues covered.

Follow-up interviews encompassed sessions with IT staff in which there were wider-ranging discussions of IS strategy, the management structure in the company, and decision-making regarding IS investment. Interviews were conducted in English, although certain sections of the questionnaire were translated into Arabic. The initial interviews took approximately two hours, excluding follow-up phone calls and emails.

IV. CASE STUDY FINDINGS

This section applies process mapping and systems profiling to the current systems portfolio at WOC. The business activities of WOC can be grouped into six top-level business processes, each of which has a number of sub-processes. These processes are briefly outlined below, along with the information systems that currently support these process activities. This is depicted in Figure 2.

1. Exploration: The company carries out seismic mapping by exploding dynamite and measuring the way in which the resulting seismic waves travel through the underground formations to detect crude oil, which when found, is subsequently drilled for, and pumped to the surface via oil wells into storage tanks. The exploration process in the company is divided into two sub-processes: Planning, and Oil Production. The Oil Production sub-process is automated with an in-house information system developed in COBOL. This records and manages all data associated with exploration of crude oil from setting up oil rigs to pumping the oil into the reservoirs, including the volume, grade, and chemical/sulphur content of the extracted crude oil. The Planning sub-process, however, is still manual in part, with some use of the Excel spreadsheet package for the recording and management of data related to seismic mapping of prospective exploration areas, and other related information.

2. Forecasting: This process has three sub-processes, all of which are partially automated with a specially designed

Microsoft Access and Excel package. The three sub-processes are Sales Forecasting, Price Modelling and Demand Planning. The Sales Forecasting sub-process takes current crude oil market data input into the system, and uses the current supply and demand of crude oil between WOC and its customers to determine some variables and constants, to produce managerial reports that are used in forecasting sales, modelling prices and managing the known and envisaged demand. The Price Modelling function manages data on different chemical compositions of the crude oil, and generates reports on the optimal price for different grades of WOC crude oil in their reservoirs. The Demand Planning function aims to improve the accuracy of revenue forecasts, align inventory levels with peaks and troughs in demand, and enhance profitability for a given channel or product. The system uses data from customer standing orders and planned sales orders, among other statistics, to generate income forecasts and align inventory levels with forecast demand. This system does not completely automate these sub-processes; there are still some supporting manual activities to obtain and enter data into the system.

3. Financial Management: The process can be subdivided into two main sub-processes: Financial Accounting, which manages the balance sheet as well as profit and loss account, different journals used, and other financial reports; and Management Accounting, which handles the financial costing activities, and generates appropriate prices for sales of crude oil. Both these sub-processes are automated using the Financials and Controlling (FICO) module of SAP. Real-time and reliable managerial reports are available from the system, and any new format of financial reporting needed can be customised and generated from the system.

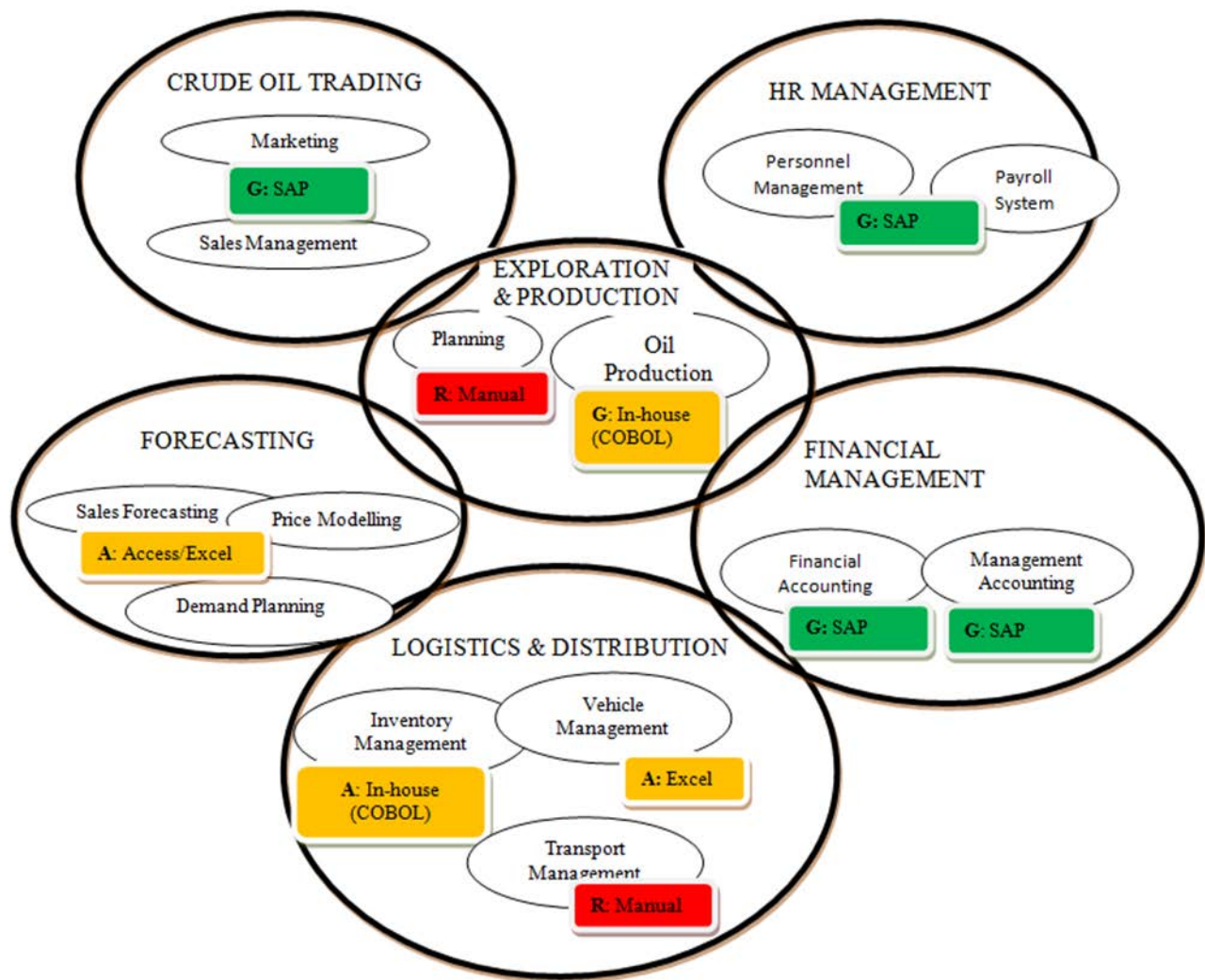
4. Logistics and Distribution: The Transportation Management sub-process is partly undertaken by third party logistics companies transporting crude oil to the customer, and supporting activities in-house are mainly manual. However, Inventory Management is automated via an in-house system developed in COBOL, and the Vehicle Management sub-process is supported by a bespoke system developed in Excel to manage the fleets of vehicles owned by WOC.

5. Crude Oil Trading: Both sub-processes of Sales Management and Marketing Management use the Sales and Distribution (SD) module of SAP. Sales Management involves raising pro forma orders and invoices, and tracking them through the sales order processing cycle. The Marketing Management sub-process encompasses customer relationship management, prospecting for new customers and recording management details of all marketing activities. The system contains up-to-date records of customers and their business activities with the company. The records are used to respond to all customer queries.

6. Human Resource Management: This is centrally managed and automated using the SAP HR module. There are two main sub-processes: Personnel Management and Payroll Management. SAP handles the creation and management of new employee records on the system, and all their professional relationships with the company, including

leave management, bonuses, and allowances. It handles the periodic assessment of staff activity and manages the payment of salaries and wages to regular and casual staff.

The top managerial reports are centrally available via the SAP system and the in-house systems developed in COBOL. Senior management rely on the reports from these different



Notes: G (Green) = indicates a system that is effective.
 A (amber) = indicates a system that may need replacement.
 R (red) = indicates a system that is defective and need replacing.

Figure 2. Main Business Processes and IS profiling at Waha Oil Company

The information system strategy adopted at WOC is a proportionate blend of an in-house/bespoke approach using COBOL and Microsoft Excel and Access, combined with a phased implementation of SAP. The choice of SAP in 2006/7 was largely pragmatic. The company’s senior management had elected to move in phases to a mainstream integrated software package, and SAP were the first such company to set up a sales and support office in Libya. There were no realistic alternatives at that time, although Oracle, another major ERP vendor, have since opened a sales office in Tripoli also.

information systems to derive a complete overview of business activities at any one point in time. This illustrates the issues that still prevail regarding systems integration. The Financial Management, the Crude Oil Trading, and Human Resource Management processes are efficiently integrated around SAP R/3 technology. These SAP modules are not well integrated with the in-house bespoke systems. However, information from the Access/Excel systems that support the Forecasting process and Logistics and Distribution processes is exported in comma delimited files (.csv format) and imported into the SAP R/3 system via data ports.

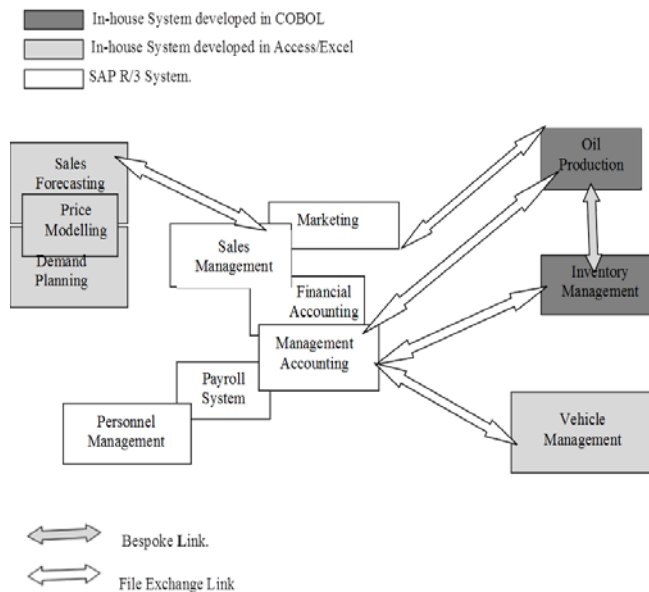


Figure 3. Systems Interfaces at Waha Oil Company

In similar vein, data is exported in a .csv format from SAP R/3 and imported for use in the Access/Excel and COBOL systems. These comma delimited files are used with human intervention to bridge the integration gap that exists between these three different groups of systems based on different technologies (see Figure 3).

V. ANALYSIS AND CONCLUSION

The current information system strategy at WOC was adopted in 2007 to support the company’s business strategy of expansion and increase in bottom-line profits. The development of this strategy was largely a top-down process undertaken by senior management, which envisaged a stage by stage implementation of the SAP ERP system, and a gradual phasing out and replacement of existing in-house developed systems. The initial focus was to be on the core sales order processing cycle and back-end financial systems, where problems of inconsistent data were paramount, and which were affecting customer service levels. In that year, an IS strategy committee was set-up by the senior management of the company, comprising members chosen from across all company departments. The committee focused on aligning the company’s business processes with the newly adopted IS strategy. Since then, over the past seven years, there has been a gradual phasing-out of the in-house systems and their replacement with appropriate SAP modules. The in-house systems that remain continue to meet the basic needs of the sub-processes that they are supporting. Benefits of this hybrid strategy can be generally grouped as follows:

1. Improved turn-around time in business activities. A typical example given by a respondent was the elimination of comparative checks of reports between the ones generated by the Financial Management process and Crude Oil Trading process. The fact that both processes are managed

by a single information system (SAP R/3) has eliminated the need for comparison of reports generated by the two processes.

2. Improved availability of key business information. Once the information from other systems is exported and imported into SAP, management reports become available on a real-time basis: for example, analysis of forecast sales vs. actual sales for a specific period. Senior executives now use systems outputs to support key business management decisions. Reports on the weekly oil production and weekly oil dispatched to customers are used as a guide on how to take decisions on maintenance and the management of the oil storage infrastructure of the company.

3. There is an increased centralisation of data maintenance in the company, which has reduced the human error that results from double or triple entry of data into different systems, and the importing and exporting of information across different systems. This has increased the reliability of corporate information and reporting, both within and outside of the company. For example, the confirmed orders, sales made, invoices generated, amounts involved, tax deducted, discount given, and all the financial stages in the Crude Oil Trading process can be accessed by authorised users in the Financial Management department, because business transactions in both the Crude Oil Trading and Financial Management processes are executed within the SAP system. This reduces time wastage that comes with reconciliation of financial activities, when data is maintained in more than one system. Questionnaire respondents confirmed a significant reduction in data error rate from in excess of 10% prior to the adoption of SAP, to current levels of around 2%.

The current plan is to complete the replacement of all legacy systems by 2017, whilst at the same time increasing the e-business capabilities of the company, exploiting the web capabilities of the SAP modules. The progress made to date in this 10-year implementation plan for SAP can be viewed in terms of the three dimensions of people, process and technology noted above in Section 2. In terms of the advance in people skills and capabilities, there has been a continuous expansion in the number of users of the SAP modules in the organisation, and the general users’ acceptance of the system appears to have contributed significantly to its successful implementation. This was engendered by a clear commitment by senior management to provide staff training, allied to their all-inclusive approach to systems implementation, whereby staff at all levels of the company were included in the process. The one area where there are arguably some significant staff issues is in the IT department, where the work-load of the current support staff is very high. There are only two staff responsible for supporting all the information systems of the company, including the legacy COBOL and Access/Excel systems.

This issue should be resolved to a degree once these legacy systems are phased out.

The new SAP technology is a strategically sound product. The centralisation of data and information provided by the SAP R/3 suite is a major benefit and its ease of use, enhancement capability, and enterprise-wide functionality has made it highly relevant to the day-to-day running of the company's business activities. Its future strategic position among the information systems in the company is guaranteed, because of the company's dependence on it, and upgrade plans for it: it is the central element of the company's IS strategy. Nevertheless, the phasing-out of the key COBOL systems may require some configuration of new SAP replacement modules. These legacy systems are used for managing oil production and management of inventory, which are crucial to the smooth running of the company's supply chain, so their replacement will need to be carefully planned and managed. They currently support business activities to an acceptable level, but their lack of flexibility and associated upgrade problems justify their replacement. The replacement of the Excel-based vehicle management system and the Access/Excel forecasting systems are less problematic, but will also require appropriate testing and configuration of additional SAP modules.

The implementation of the new SAP modules has been co-ordinated with change management with regard to improvement of people capabilities, knowledge and training. A key to success at the company is that there has been a steady improvement in processes rather than any radical change, and change has been centred on the sales order processing and production backbone of the company. The change in the three areas of people, technology and process improvement has been balanced and focussed. This was underpinned by a staged implementation of new SAP modules, that has already spanned seven years, and should complete in a further three years. This has allowed careful management and firm cost control of the phased migration to the SAP product suite.

REFERENCES

- [1] C. Koch, "The ABCs of ERP", CIO Magazine, Dec 22nd, 1999.
- [2] A. Hawari and R. Heeks, "Explaining ERP failure in a developing country: a Jordanian case study," *Journal of Enterprise Information Management*, vol. 23, no. 2, 2010, pp. 135-160.
- [3] M. Moohebat, M. Jazi, and A. Aseni, "Evaluation of ERP implementation at the Esfahan Steele Company," *International Journal of Business and Management*, vol. 6, no. 5, 2011, pp. 236-250.
- [4] P. J. DiMaggio and E. Hargittai "From the Digital Divide to Digital Inequality", Working Paper 19, 2001, Centre for Arts and Cultural Policy Studies, Woodrow Wilson School, Princeton University.
- [5] M. Warschauer, "Dissecting the Digital Divide: a case study in Egypt," *The Information Society*, vol. 19, 2003, pp. 297-304.
- [6] R. Gomez and S. Pather, "ICT evaluation: are we asking the right questions?", *Electronic Journal of Information Systems in Developing Countries (EJISDC)*, vol. 50, no. 5, 2012, pp. 1-14.
- [7] S. Batchelor, S. Evangelista, S. Hearn, M. Pierce, S. Sugden, and M. Webb, *ICT for development: contributing to the millennium development goals – lessons learnt from seventeen infoDev projects*, Washington DC: World Bank, 2003.
- [8] R. Heeks, "Information systems and developing countries: failure, success, and local improvisations", *Journal of Information Society*, vol. 18, no. 2, 2002, pp. 101-112.
- [9] C. Soh and S Kien Sia, "An institutional perspective on sources of ERP package-organisation misalignments", *Journal of Strategic Information Systems*, vol. 13, 2004, pp. 375-397.
- [10] Department of Trade and Industry, *Business in the information age: the international benchmarking study*, 2004, London: Booz Allen Hamilton.
- [11] P. Harmon, *Process Maturity Models*, BP Trends, 2009. available: <http://www.bptrends.com/bpt/wp-content/publicationfiles/spotlight_051909.pdf>. [Retrieved 20 March 2014].
- [12] M. Wynn and O. Olubanjo, "Demand-supply chain management: systems implications in an SME packaging business in the UK", *International Journal of Manufacturing Research*, vol. 7, no. 2, 2012, pp. 198-212.
- [13] K. M. Eisenhardt, "Making fast strategic decisions in high-velocity environments", *Academy of Management Journal*, vol. 32, no. 3, 1989, pp. 543-576.
- [14] M. Easterby-Smith, R. Thorpe and P. Jackson, *Management Research*, 2012, Sage Publications.
- [15] A. Bryman, A. and E. Bell, *Business Research Methods*, 3rd edition, Oxford: Oxford University Press, 2011.
- [16] M. Saunders, P. Lewis and A. Thornhill, *Research methods for business students*, 5th edn., 2009, England: Pearson Education Limited.
- [17] J. Hartley, "Case studies in organisational research", in C. Cassell and G. Symon (eds), *Essential guide to qualitative methods in organizational research*, 2004, London: Sage. pp 323-333.
- [18] R. K. Yin, *Applications of Case Study Research*. 3rd edn., 2012, London: SAGE Publications, Inc
- [19] H. Simonds, *Case Study Research in Practice*, Sage Publications, 2009.

The Key Contributions of the Operations Management and Information Systems Disciplines to Business Process Management

Philippe Marchildon and Pierre Hadaya
Department of Management and Technology
ESG-UQÀM
Montréal, Canada
e-mail: marchildon.philippe@courrier.uqam.ca
hadaya.pierre@uqam.ca

Abstract—Based on a narrative review, this paper synthesizes the main contributions of the operations management and the information systems disciplines to the business process management literature. Our findings show that the operations management discipline has been the main contributor to the topics of business process definition, business process standardization, business process outsourcing/offshoring, Six Sigma, and business process management theories while the information systems discipline has been the main contributor to the topics of business process reengineering, the role of information technology, and the business process management nomological network.

Keywords—business process management; information systems discipline; operations management discipline

I. INTRODUCTION

The study of business processes has been a long-standing concern for members of academia and practitioners [1]. A business process can be defined as a lateral or horizontal organizational form that encapsulates the interdependence of tasks, role, people, departments and functions required to provide a customer (either internal or external) with a product or service, through the transformation of inputs into outputs [2][3][4][5]. The term customer refers here to both external consumers of the organization and internal recipients at linkage point between processes, as output from upstream processes become the input of subsequent processes [3][4]. This wealth of attention on business processes has fostered a vast literature centered on business process management (BPM). BPM can be defined as a strategy-driven organizational initiative to improve and (re)-design business processes to achieve competitive advantage through changes in the relationships between management, information, technology, organizational structure, and people [6][7].

Two disciplines have mainly contributed the BPM literature: the operations management (OM) and information systems (IS) disciplines. The OM discipline is concerned with overseeing, designing, and controlling the process of production and redesigning business operations in the production of goods or services by considering the acquisition, development, and utilization of resources. The IS discipline examine the phenomena that emerge when

technology and peoples interact. While both disciplines have brought significant contributions to BPM, their respective efforts are most often conducted in silo and are rarely integrated into a common whole. Consequently, we still lack a comprehensive understanding of the current state of knowledge on BPM. The objective of this paper is thus to synthesize the main contributions of the OM and IS disciplines in order to comprehensively detail the state of knowledge on BPM.

The findings exposed in this paper are based on a narrative review. A narrative review provides a verbal summary of previously published research on a particular topic of interest by either focusing on related concepts and theories, research methods or research outcomes [8]. Narratives reviews “serve a scientific field by providing a much-needed bridge between the vast and scattered assortment of articles on a topic and the reader who does not have the time or resources to track them down [9, p. 311]”. In addition, narrative reviews can serve as an appropriate starting point for future inquiries and research developments [10].

The rest of the paper is organized as follows. Section 2 presents a brief history of the BPM literature to demonstrate that today’s understanding of business processes and their improvement highly rest on the intertwined findings of the OM and IS disciplines. Then, based on this understanding, we use a narrative approach in Section 3 to detail the main contributions of the OM and IS disciplines along the four main sub-streams of the BPM literature: business process standardization (BPS), business process reengineering (BPR), Six Sigma, and theorizing efforts. The paper concludes with a presentation of research limits and future research avenues.

II. A BRIEF HISTORY OF BUSINESS PROCESS MANAGEMENT

Initially, the study of business processes emerged as a central element of total quality management (TQM) [3][7]. The history starts with the seminal work of quality management proponents such as Ishikawa [11], Deming [12], and Juran [13]. In essence, these works focused on “the creation of an organizational system that fosters cooperation and learning for facilitating the implementation of process

management practices, which in turn, leads to continuous improvement of processes, products and services, and to employee fulfillment, both of which are critical to customer satisfaction, and, ultimately to firm survival [14, p. 473]". Subsequent writings by Davenport and Short [15] and Hammer [16] highlighting the necessity to focus on business processes reengineering (BPR) reinforced and broaden the initial interest on the subject [7]. It was also at this point in time that the interdependent relationship between IS and BPM was explicitly acknowledged [15] enticing the OM and IS communities to work together to improve our understanding of BPM. This is certainly exemplified by the adjacent publication of special issues on the topic of TQM in the Academy of Management Review in June 1994 and the Journal of Management Information Systems in 1995, two top journals in their respective discipline.

Today, with the increasing emphasis placed on integrating business Web sites with backend legacy and enterprise systems, the management of business processes remains an important topic in the IS discipline [17] while the need for ambidextrous organizations puts BPM to the forefront in the OM and management discipline [3]. Hence, even though the initial idea of BPM emerged from the OM discipline [15], today's understanding of business processes and their improvement highly rests on the intertwined findings of the OM and IS disciplines. Two reasons explain why BPM has been predominantly researched in these two disciplines. First, researchers within both the OM and IS communities have long recognized the systemic nature of the firm and the need for a holistic approach in its management [18][19]. Hence, studying business processes, which span across intra-organizational and in certain circumstances inter-organizational boundaries, is important in both disciplines. Second, because business processes span across internal and external organizational boundaries and because one of the key purposes of information technology (IT) is to reduce coordination cost across organizational entities, both disciplines have recognized the complementary if not symbiotic nature of business processes and IT [15][20][21], creating a state where both communities have mutually reinforced each other's work and interest on BPM.

III. A DETAILED NARRATIVE OF BUSINESS PROCESS MANAGEMENT AND KEY CONTRIBUTIONS

Having briefly exposed the history of BPM and the importance of both the OM and IS disciplines to fully comprehend the improvement of business processes, we now examine the BPM literature in greater detail to show the contributions of both disciplines along the four critical sub-streams of the BPM literature. To do so, we first show that the BPM literature initially evolved into two separate sub-streams, BPS and BPR while a third sub-stream entitled Six Sigma that reconciles diverging views from the BPS and BPR sub-stream appeared more recently. Next, we detail the contributions within the fourth sub-stream that has focused on theorizing efforts to provide BPM explanations and to develop a BPM nomological network in an attempt to alleviate the previous three sub-stream's shortcomings. We

conclude by synthesizing the key contribution of both disciplines to BPM.

A. *The business process standardization, business process reengineering and Six Sigma sub-streams*

Various programs like TQM, ISO 9000, the Malcom Baldrige Award, Six Sigma and BPR have been developed to help managers improve their business processes [3][14][22]. However, despite the fact these programs share several similarities (i.e., they all aim to improve business processes), they also differ in scope and approach [3] as they differ significantly in the magnitude of change sought-after to improve business processes [3][23]. For instance, TQM, ISO 9000, and the Malcom Baldrige Award programs have been depicted as programs seeking incremental changes [14][23][24] whereas BPR programs have been described as programs aiming for radical changes [6][16][25]. This major difference may be explained by the fact that proponents of both types of programs have different backgrounds. Indeed, TQM, ISO 9000, and the Malcolm Baldrige Award advocates relying on their vast experience with statistical process control continue to argue for incremental changes while BPR advocates relying on their IT implementation experience prone for radical changes [16][18]. Consequently, this divergence in scope and approach has led to the emergence of two key sub-streams in the BPM literature. The first, BPS, is mainly addressed in works from the OM discipline while the second, BPR is mainly addressed in works from the IS discipline. The BPS sub-stream has mainly focused on the standardization of business processes and process outsourcing/offshoring (BPO) while the BPR sub-stream has mainly addressed the reengineering of business processes and the role of IT in BPM. More recently, a third sub-stream focusing on Six Sigma has emerged. With its roots in the OM discipline, this sub-stream reconciles findings from the BPS and BPR sub streams.

1) *Business Process Standardization*

The cumulative and extensive work of TQM and other OM programs (e.g., ISO 9000, the Malcom Baldrige Award) has led to the identification of three key principles for business improvement, namely focus on customers and stakeholders, participation and teamwork throughout the organization, and focus on continuous improvement and learning [26], as well as the creation of three key BPM components: (1) process activity and flow standards, (2) process performance standards and (3) process management standards [4]. These principles and components have enabled a more efficient approach to improve business processes while simultaneously allowing for the emergence of BPS [27]. BPS can be defined as the degree to which work rules, policies and operating procedures in an organization, as established by consensus and approved by a recognized body (e.g., government agency, industrial consortia) [28], are formalized and followed [29]. BPS offers several important benefits to organizations [27]. Within an intra-organizational context, BPS facilitates communications on business operations, enables smooth handoff across process boundaries, and makes possible comparative measures of performance. Likewise, within an inter-organizational

context, BPS makes commerce easier by improving communication, enabling more efficient handoffs and allowing performance benchmarking [4]. As such, BPS has been shown to foster economies of scale, organizational learning, and overall organizational effectiveness [27].

Furthermore, when pushed to the extreme, the idea of BPS has also led some protagonists from operations management to believe that business processes could be outsourced/offshored in order to allow firms to reap further benefits by concentrating their efforts on their core competencies [4][30][31]. BPO can be defined as the delegation of one or more business processes to an external provider, whether onshore or offshore (Mani et al., 2010). Recent findings emanating from the OM and IS disciplines, however, suggest that reaping benefits from this approach is not as straightforward as previously expected [27][32] and that these benefits may only be temporary. Thus, organizations seeking to gain a sustainable competitive advantage should proceed carefully in embracing BPS and/or BPO practices [33][34].

2) *Business Process Reengineering*

BPR can be defined as an approach “for initiating and managing “radical” changes in business processes [35, p. 32]”. Hence, with the help of statistical and quantitative analysis, BPR advocates aim to fundamentally rethink and redesign business processes in order to obtain dramatic and sustainable improvements in contemporary measures of performance (e.g., quality, cost, service, lead time, outcomes, flexibility, innovation) [25][26]. On their quest for radical changes, BPR advocates have directed their attention away from business processes themselves and started to look for solutions that could significantly alter them. This resulted on a strong emphasis on IT due to its ability to reduce coordination costs across internal and external organizational boundaries [15]. Accordingly, BPR advocates are mostly IS researchers. They have proposed five distinctive steps, one of which is dedicated to IS, to help managers reengineer business processes [15]. As a first step, managers should develop a business vision and define clear process objectives. Second, managers should identify the process/processes to be redesigned. Third, managers should develop an understanding of existing processes and measure them. Fourth, managers should identify key IT levers and fifth, managers should design and build a prototype of the improved business process/processes. By examining extensively the key role of IT in improving business processes, BPR advocates were able to determine that IT contributes to the reengineer of business processes in two ways. First, by forming an organization’s information backbone that spans across functional level and enables easier communication. Second, by providing capabilities that support key BPR activities, such as modeling, optimizing and validating [6][7][17]. Hence, IT can be seen as both an enabler [20][25][36] and as a facilitator [37][38] of BPR. More precisely, IT plays an enabler’s role when it is used as a fundamental component of an improved process whereas IT plays a facilitator’s role when it supports the process improvement process without being included as a fundamental component of the final solution (i.e., the new or

improved process doesn’t require IT). For example, improving the customer payment process may rest on the added capability of information technologies that allows for automatic payment between firms (i.e., IT enabler’s role) whereas improving a product assembly process could be done through the use of statistical software in order to create a new optimal assembly sequence. In this latter example, the assembly process remains IT free but IT played a key role in improving the process (i.e., IT facilitator’s role). Evidently, both roles are not mutually exclusive and IT can, in many cases, play both roles simultaneously.

Despite showing the great power of IT, these efforts have also highlighted the limits of technology. Indeed, these efforts have demonstrated that IT should not be a panacea to organizational process improvement problems but rather be considered as a part of a broader approach. That is, implementing IT just for the sake of it is not going to improve a business process. Having IT in mind as either an enabler, facilitator or both, organizations should aim to remodel their processes, in a way that the new processes developed answer business needs [20][38]. Furthermore, BPR efforts have also highlighted that IT can be a barrier to business process improvement. For instance, a firm lacking interoperability between its data from different information systems was unable to implement an improved version of its replenishment processes because its selling systems could not be readily integrated with ordering and logistics systems [36]. Taken as whole, these complementary findings indicate that anchoring BPR or BPM on IT alone is not enough to provide a sustainable competitive advantage [20][38] while the long term consequences of IT have to be considered since today’s solution can become tomorrow’s problems.

3) *Six Sigma*

Despite being mainly treated in two different disciplines, and characterizing business process change in a dichotomous manner where the relationship between incremental and radical changes is mutually exclusive, the BPS and BPR sub-streams now seem to be converging. Indeed, recent characterizations of business process change in the BPM literature now follow a less strict standpoint and depict business process changes on a continuum ranging from incremental to radical changes, making the simultaneous pursuit of both types of changes possible [6][7][37]. This reconciliation between BPS and BPR advocates highlights the similarities between the two programs. That is, both BPS and BPR rest on the common purpose of transforming business processes by measuring, improving, and rationalizing each individual process as well as the handoffs between the different processes [3][7][21]. This convergence also highlights that the improvement of business process is grounded in three main common practices: mapping processes, improving processes, and adhering to systems of improved processes [3].

The reconciliation between BPS and BPR is also at the hearth of the emergence of new sub-stream on Six Sigma and may explain why organizations such as 3M, Ford, Honeywell and American Express already pursuing TQM and BPR programs were able to reach further benefit by adopting Six Sigma [24]. Indeed, because Six Sigma allows

organizations to be ambidextrous, that is to support simultaneously the need for exploration and control, this innovative program combines the advantages of BPR and BPS while minimizing their respective shortcomings [3][24]. Six Sigma can be defined as a “project-driven management approach to improve the organization’s products, services, and processes by continually reducing defects in the organization [39, p. 1]”. As such, the main difference between BPS, BPR and Six Sigma rests in how business improvement tools/techniques are implemented in the organization, rather than in the underlying philosophy or the tools/technique employed to improve business processes [24]. More precisely, Six Sigma differs from BPR as it places more emphasis on data driven decisions rather than on statistics and quantitative analysis [26]. On the other hand, Six Sigma differs from BPS on the following aspects. First, it provides a more structured and rigorous training development program for managers. Second, the business process and its improvement is owned by a single “champion” in Six Sigma rather than by a multitude of worker in BPS. Third, Six Sigma is cross-functional and looks for verifiable return on investment whereas BPS is a process based methodology that lightly focuses on financial accountability [26]. Thus, besides providing a platform to allow for both incremental and radical changes, Six Sigma also suggests that an integrative framework of BPM is coming of age. It is important to note however that, to this day, the topic of Six Sigma has mainly been discussed in the OM disciplines while being addressed in only a very limited number of IS studies [40].

B. The Sub-Stream on Theorizing Efforts

Research within the three previous sub-streams has fostered our knowledge on business process and BPM. However, although essential, these efforts remain insufficient to provide a comprehensive understanding of business process improvement. This is certainly exemplified by the fact that numerous if not the majority of organizations adopting one or many of the BPM programs mentioned above actually fail to reach expected benefits [1][6][41][42]. This phenomenon has created a productivity paradox with some organizations reaping significant benefits from BPM while others actually losing money. This issue is further exacerbated by the limited number of empirical research conducted to assess the effectiveness of BPM programs which has resulted in a state where BPM programs tailored to improve business processes are usually developed and adopted on the basis of anecdotal evidence rather than scientific knowledge [3][7][14]. Put differently, these previous observations indicate an evident lack of BPM theorizing [1][14][22]. In accordance with this assertion, several authors have observed that the vast majority of the studies on business process and BPM remains to this day highly prescriptive in nature and thus fails to highlight the underlying mechanisms behind the various programs developed and their respective limits [3][22].

Recognizing the need for theory, the OM and IS disciplines have conjointly begun to theorize on BPM. To do so, they have adopted various approaches: identifying

BPM/IT critical success factors [1][41], identifying BPM antecedents [6], linking BPM with existing management theory [22], building BPM theory by using grounded theory [24] and censuring current methodologies, techniques and tools [5] in an effort to resolve the issue. Representative findings from these theorizing efforts are summarized in Table 1.

Four broad assertions can be gleaned from these theorizing efforts. First, BPM builds from knowledge rooted in multiple disciplines including management as strategy, organizational behavior and psychology, industrial economics and purchasing, innovation, organization design and human resources, sociotechnical design, quality and industrial engineering, marketing and finance [3][7]. Considering that the idea was to foster a holistic approach to organization management, it is not surprising that business process theorizing efforts have drawn from multiple disciplines that, altogether, allow for the required 360 degrees view of an organization. Second, theorizing efforts aiming to explain the impact of BPM on organizational performance position the construct of business process management in a complex and dense nomological network [3][6][7][35]. A clear insight stemming from these proposed nomological networks is that researchers agree on the start and end point of BPM. Specifically, proposed BPM nomological networks typically build on the premise that BPM initiatives should be triggered by a strategic vision and aim for customer focused outcomes [6][7][14][22].

TABLE I. REPRESENTATIVE INSIGHTS FROM BPM THEORIZING EFFORTS

Authors (discipline)	Insights
[1] (IS)	Development of a theoretically rooted framework identifying business process management critical success factors.
[3] (OM)	Development of a theoretical contingency approach to business process management based on the constructs of process management, technological innovation, organizational environment, organizational form and organizational adaptation.
[5] (IS)	Identification of a comprehensive list of methodologies, techniques, and tools supporting business process management.
[6] (IS)	Empirical validation of Kettinger and Grover’s [7] theoretical framework highlighting the validity of the framework.
[7] (IS)	Development of a multilevel theoretical framework of business process change management including 10 elements and their relationships
[14] (OM)	Identification of the concepts and their relationships underlying the Deming Management Method.
[18] (OM)	Characterization of business processes along the dimensions of work processes, behavioral processes, and change processes.
[22] (OM)	Development of a theoretical framework highlighting the similarities between the total quality and management literature based on the main dimensions of the Baldrige Award.
[24] (OM)	Development of a Six Sigma framework including 5 elements and their relationships based on the identification of an underlying theory on Six Sigma.
[35] (IS)	Development of a taxonomy of BPR strategies based on a process alignment model comprising four lenses: process, strategy, information systems, and change management.

Third, contrary to early prescriptive attempts that prone the universality of BPM programs, current efforts clearly highlight that BPM programs are context dependent. The identified contingency factors generally include elements both internal and external to the organization [3][6][7]. Internal factors can be categorized in terms of organizational structure, management, information and technology, people, and business processes [6][7] while external factors usually refer to environmental conditions (e.g., economic conditions, industry competitiveness, innovations) [3][7]. As such, a careful reflection must be made before adopting one or many BPM programs [3]. Lastly, although they differ slightly, the contribution of the OM and IS disciplines to BPM theorizing efforts are complementary. Indeed, members of the OM discipline have mainly aimed to identify and define BPM underlying theories whereas members of the IS discipline have mainly aimed to define BPM’s nomological network.

C. Synthetizing the key contribution of the operation management and IS discipline

Having described the four sub-streams of the BPM literature, their respective key research topics and highlighted the role of the OM and IS disciplines in regards of each of these topics, we can now compare each discipline’s contribution towards the improvement of business processes. Table 2 synthetizes the results of this comparison.

TABLE II. THE DISTINCTIVE CONTRIBUTION OF THE OPERATIONS MANAGEMENT AND IS DISCIPLINES ON BUSINESS PROCESS KNOWLEDGE

Topic	OM	IS	Dominant discipline
Definition of processes	[4][1][18][21]	[2]	OM
Definition of business process management	[1][18][43][44][45][46]	[6][7]	Equivalent
Business process standardization	[3][4][24][27, 28]	[2][34]	OM
Business process outsourcing/offshoring	[4][30][31][32][33]	[37][38][47]	OM
Business process reengineering	[15][41]	[17][20][23][25][36][48]	IS
The role of IT	[15][43]	[5][6][7][17][20][35][36]	IS
Six Sigma	[24][26][45][46]	[39]	OM
BPM theories	[3][14][22][24][45][46]	[6][7]	OM
BPM nomological network	[3][24]	[6][7][35][49]	IS

Keeping in mind the results described above, one can see that the OM discipline played a prominent role in the topics of business processes definition, BPS, BPO, Six Sigma, and BPM theories. On the other hand, the IS discipline played a prevalent role in the topics of BPR, the role of IT, and BPM nomological network. Finally, both disciplines had an equal contribution in the business process management definition topic.

IV. CONCLUSION

We set out to identify the key contributions of the OM and IS disciplines to BPM. Our findings suggest that our knowledge of business processes and their improvement rests on the intertwined work of the OM and IS disciplines as neither discipline has comprehensively addressed each key BPM topics. While we have used a narrative review to provide a preliminary portrait of the OM and IS disciplines’ contributions to BPM and to show that each discipline seems to focus on different BPM topics, we have yet to thoroughly assess the quality of the sources composing our narrative review. Future research could address this issue through citation analysis and/or expert discussions, which would provide a more objective assessment of each discipline’s key contributions to BPM.

REFERENCES

- [1] P. Trkman, “the critical success factors of business process management”, *International Journal of Information Management*, vol. 30, no. 2, 2010, pp. 125-134.
- [2] A. Basu and R. W. Blanning, “Synthesis and Decomposition of Processes in Organizations”, *Information Systems Research*, vol. 14, no. 4, 2003, pp. 337-355.
- [3] M. J. Benner and M. L. Tushman, “Exploitation, exploration, and process management: The productivity dilemma revisited”, *Academy of Management Review*, vol. 28, no. 2, 2003, pp. 238-256.
- [4] T. H. Davenport, “The Coming Commoditization of Processes”, *Harvard Business Review*, vol. 83, no. 6, 2005, pp. 100-108.
- [5] W. J. Kettinger, J. T. C. Teng, and S. Guha, “Business Process Change: A Study of Methodologies, Techniques, and Tools”, *MIS Quarterly*, vol. 21, no. 1, 1997, pp. 55-80.
- [6] S. Guha, V. Grover, W. J. Kettinger, and J. T. C. Teng, “Business Process Change and Organizational Performance: Exploring an Antecedent Model”, *Journal of Management Information Systems*, vol. 14, no. 1, pp. 119-154, 1997.
- [7] W. J. Kettinger and V. Grover, “Special Section: Toward a Theory of Business Process Change Management”, *Journal of Management Information Systems*, vol. 12, no. 1, 1995, pp. 9-30.
- [8] G. Paré, M.-C. Trudel, M. Jaana, and S. Kitsiou, “Synthesizing Information Systems Knowledge: A Taxonomy of Review Types”, *Cahier du GReSI, HEC Montréal*. 2013.
- [9] R. F. Baumeister and M. R. Leary, “Writing Narrative Literature Reviews”, *Review of General Psychology*, vol 1, no. 3, 1997, pp. 311–320.
- [10] P. Cronin, F. Ryan, and M. Coughlan, “Undertaking a Literature Review: A Step-by-Step Approach”, *British Journal of Nursing*, vol 17, no. 1, 2008, pp. 38–43.
- [11] K. Ishikawa, “What is total quality control? The Japanese way”, Englewood Cliffs, NJ, Prentice-Hall, 1985.
- [12] E. W. Deming, “Out of Crisis”, Cambridge, MA, MIT Press, 1986.
- [13] J. Juran, “Juran on leadership for quality”, New York, NY, Free Press, 1989.
- [14] J. C. Anderson, M. Rungtusanatham, and R. G. Schroeder “A theory of quality management underlying the Deming management method”, *Academy of Management Review*, vol. 19, no. 3, 1994, pp. 472-509.
- [15] T. H. Davenport and J. E. Short, “The New Industrial Engineering: Information Technology and Business Process

- Redesign”, *Sloan Management Review*, vol. 31, no. 4, 1990, pp. 11-27.
- [16] M. Hammer, “Reengineering work: don’t automate, obliterate”, *Harvard Business Review*, vol. 68, no. 4, 1990, pp. 104-112.
- [17] M. Attaran, “Exploring the relationship between information technology and business process reengineering”, *Information & Management*, vol. 41, no. 5, 2004 pp. 585-596.
- [18] D. A. Garvin, “The Processes of Organization and Management”, *Sloan Management Review*, vol. 39, no. 4, 1998, pp. 33-50.
- [19] D. Miller, “Toward a new contingency approach: The search fro organizational gestalten”, *Journal of Management Studies*, vol. 18, no. 1, 1981, pp. 1-26.
- [20] T. H. Clark and D. B. Stoddard, “Interorganizational Business Process Redesign: Merging Technological and Process Innovation”, *Journal of Management Information Systems*, vol. 13, no. 2, 1996, pp. 9-28.
- [21] D. A. Garvin, “Leveraging Processes for Strategic Advantage”, *Harvard Business Review*, vol. 73, no. 5, 1995, pp. 77-90.
- [22] J. W. Dean and D. E. Bowen, “Management theory and total quality: Improving tesearch and practice through theory development”, *Academy of Management Review*, vol. 19, no. 3, 1994, pp. 392-418.
- [23] P. O’Neill and A. S. Sohal, “ Business Process Reengineering A review of rececent literature”, *Technovation*, vol. 19, no. 9, 1999, pp. 571-581.
- [24] R. G. Schroeder, K. Linderman, C. Liedtke, and A. S. Choo, “Six Sigma: Definition and underlying theory”, *Journal of Operations Management*, vol. 26, no. 4, 2008, pp. 536-554.
- [25] A. Gunasekaran and B. Nath, “The role of information technology in business process reengineering”, *International Journal of Production Economics*, vol. 50, no. 2, 1997, pp. 91-104.
- [26] D. Drake, J. S. Sutterfield, and C. Ngassam, “The Revolution of Six-Sigma: An Analysis of its Theory and Application”, *Academy of Information and Management Sciences Journal*, vol. 11, no. 1, 2008, pp. 29-44.
- [27] S.-W. Kwon, “Does the Standardization Process Matter? A study of Cost Effectiveness in Hospital Drug Formularies”, *Management Science*, vol. 54, no. 6, 2008, pp. 1065-1079.
- [28] B. Munstermann, A. Eckhardt, and T. Weitzel, “The performance impact of business process standardization”, *Business Process Management Journal*, vol. 16, no. 1, 2010, pp. 29-56.
- [29] D. Beimborn, N. Joachim, F. Gleisner, and A. Hackethal, “The Role of Standardization in Achieving IT Business Value”, *Proceedings of the 42nd Hawaii International Conference on System Sciences*, Big Island, Hawaii, 2009, pp. 1-10.
- [30] R. Metters, “A typology of offshoring and outsourcing in electronically transmitted services”, *Journal of Operations Management*, vol. 26, no. 2, 2008, pp. 198-211.
- [31] W. Youngdahl and K. Ramaswamy, “Offshoring knowledge and service work: A conceptual model and research agenda”, *Journal of Operation Management*, vol. 26, no. 2, 2008, pp. 212-221.
- [32] A. Stringfellow, M. B. Teagarden, and W. Nie, “Invisible cost in offshoring service work”, *Journal of Operation Management*, vol., 26, no. 2, 2008, pp. 164-179.
- [33] Y. Shi, “Today’s Solution and Tomorrow’s Problem”, *California Management Review*, vol. 49, no. 3, 2007, pp. 27-44.
- [34] R. J. Kauffman and J. Y. Tsai, “With or without you: The countervailing forces and effects of process standardization”, *Electronic Commerce Research and Applications*, vol. 9, no.4, 2010, pp. 305-322.
- [35] M. J. Earl, J. L. Sampler, and J. E. Short, “Strategies for Business Process Reengineering: Evidence from Field Studies”, *Journal of Management Information Systems*, vol. 12, no., 1, 1995, pp. 31-56.
- [36] M. Broadbent, P. Weill, and D. St.Clair, “The Implications of Information technology Infrastructure for Business Process Redesign”, *MISQuarterly*, vol. 23, no. 2, 1999, pp. 159-182.
- [37] D. Mani, A. Barua, and A. Whinston, “An Empirical Analysis of the Impact of Information Capabilities Design on Business Process Outsourcing Performance”, *MISQuarterly*, vol. 34, no. 1, 2010, pp. 39-62.
- [38] J. Whitaker, S. Mithas, and M. S. Krishnan, “Organizational Learning and Capabilities for Onshore and Offshore Business Process Outsourcing”, *Journal of Management Information Systems*, vol. 27, no. 3, 2010, pp. 11-42.
- [39] H. Y. Kwack and F. T. Anbari, “Benefits, obstacles, and future of Six Sigma approach”, *Technovation*, vol. 26, no. 5, 2006, pp. 708-715.
- [40] R. McAdam, S.-A. Hazlett, and J. Henderson, “A critical review of Six Sigma: Exploring the dichotomies”, *The International Journal of Organizational Analysis*, vol. 13, no. 2, 2005, pp. 151-174.
- [41] M. Terziovski, P. Fitzpatrick, and P. O’Neill, “Successful predictors of business process reengineering (BPR) in financial services”, *International Journal of Production Economics*, vol. 84, no. 1, 2003, pp. 35-50.
- [42] J. Karim, T. M. Somers, and A. Bhattacharjee, “The impact of ERP implementation on business process outcomes: A factor-based study”, *Journal of Management Information Systems*, vol. 24, no. 1, 2007, pp. 101-134.
- [43] W.M.P. van der Aalst, A.H.M. ter Hofstede, and M. Weske, “Business Process Managemetn: A Survey”, *Proceedings of the 1st International Conference on Business Process Management*, Eindhoven, Netherlands, 2003, pp. 1-12.
- [44] M. Weske, W.M.P. van der Aalst, and H.M.W. Verbeek, “Advances in Business Process Management”, *Data and Knowledge Engineering*, vol. 50, no. 1, 2004, pp.1-8.
- [45] J. vom Brocke and M. Rosemann “Handbook on Business Process Management 1”, Heidelberg, Germany, Springer, 2010.
- [46] J. vom Brocke and M. Rosemann “Handbook on Business Process Management 2”, Heidelberg, Germany, Springer, 2010.
- [47] H. Tanriverdi, P. Konana, and L. Ge, “The Choice of Sourcing Mechanism for Business Processes”, *Information Systems Research*, vol. 18, no. 3, 2007, pp. 280-299.
- [48] V. Grover, R. S. Jeong, W. J. Kettinger, and J. T. C. Teng, “The Implementation of Business Process Reengineering”, *Journal of Management Information Systems*, vol. 12, no. 1, 1995, pp. 109-144.
- [49] L. V. Orman, “A Model Management Approach to Business Process Reengineering”, *Journal of Management Information Systems*, vol. 15, no. 1, 1998, pp. 187-212.

Mapping the Fuzzy Semantic Model into Fuzzy Object Relational Database Model

Sabrina Jandoubi
and Nadia Yacoubi-Ayadi

National School of
Computer Sciences
University of Manouba
Tunis, Tunisia

Email: sabrine.jandoubi@gmail.com
nadia.yacoubi.ayadi@gmail.com

Afef Bahri

MIRACL Laboratory
High School of Computing
and Multimedia
University of Sfax
Sfax, Tunisia

Email: afef.bahri@gmail.com

Salem Chakhar

Portsmouth Business School and
Centre for Operational Research
and Logistics
University of Portsmouth
Portsmouth, UK

Email: salem.chakhar@port.ac.uk

Abstract—This paper discusses the mapping of the Fuzzy Semantic Model (FSM) into a Fuzzy Object Relational database Model (FuzzORM). We designed a set of mapping rules to transform all the constructs of the FSM into the FuzzORM. A prototype supporting these rules is under development over the Object-Relational Database Management System (ORDBMS) PostgreSQL. The first results of implementation are presented in this paper.

Keywords—Fuzzy database; Imperfect information; Mapping rule; Object relational database; Semantic modeling.

I. INTRODUCTION

The semantic data models are powerful conceptual modeling tools but lack effective implementation mechanisms. Thus, most of fuzzy semantic data models have been mapped and implemented through relational [1][2] or object-oriented [3][4] database models. This paper discusses the mapping of the FSM [5] into a FuzzORM. This solution permits to take advantages of both relational and object-oriented database models.

We designed a set of mapping rules to transform the constructs of the FSM into the FuzzORM. A prototype supporting these mapping rules is under development over the ORDBMS PostgreSQL. The first results of implementation are presented in this paper.

The rest of the paper is organized as follows. Section II briefly reviews the FSM. Section III details the mapping rules. Section IV presents the first implementation results. Section V discusses some related work. Section VI concludes the paper.

II. FUZZY SEMANTIC MODEL

In this section, we provide a very brief review of FSM [5].

A. Fuzzy Classes

Let E be the universe of discourse. A fuzzy entity e in E is a natural or artificial entity that one or several of its properties are fuzzy. At the extensional level, a fuzzy class K in E is a collection of fuzzy entities having some similar properties: $K = \{(e, \mu_K(e)) : e \in E \wedge \mu_K(e) > 0\}$, where $\mu_K : E \rightarrow [0, 1]$ is the membership function that maps the elements of E to the range $[0, 1]$, and $\mu_K(e)$ represents the degree of membership (d.o.m) of the fuzzy entity e in class K . At the intensional level, a fuzzy class K is defined as a set of attributes and a set of decision rules.

Figure 1 shows a FSM-based model example.

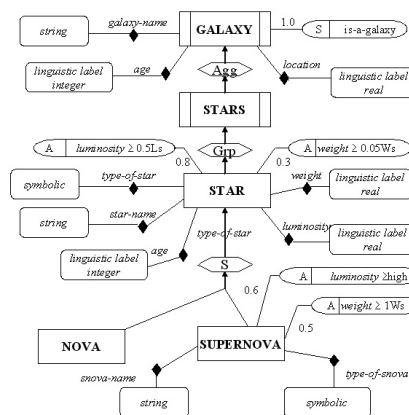


Figure 1. Example of FSM model.

For example, the class STAR in Figure 1 has five attributes (*star-name*, *type-of-star*, *age*, *luminosity* and *weight*) and two decision rules ('*luminosity* ≥ 0.5 ' and '*weight* ≥ 0.05 ').

B. Attributes

Each attribute is basically characterized by its name, data type and domain. A data type may be crisp or fuzzy. The domain of an attribute $attr$ is the set of values the attribute may take. Let $T(attr)$ denotes the domain of $attr$. Domains of fuzzy attributes are also fuzzy. For instance, the fuzzy attribute *location* associated with class GALAXY in Figure 1 has the following domain: {in, near, very near, distant, very distant}.

C. Decision Rules

Decision rules may be based on attributes or on common semantics. An attribute-based decision rule is a condition of the form $\langle attr \rangle \langle op \rangle \langle v \rangle$, where $attr$ is an attribute, op is a binary or a set operator; and $v \in T(attr)$. A semantic decision rule is a semantic phrase used to specify the members of a fuzzy class. Two decision rules from Figure 1 are: '*luminosity* = very high' and '*is-a galaxy*'. The first decision rule is based on attribute *luminosity* and associated with class STAR. The second is a semantic decision rule associated with class GALAXY. An advanced definition of decision rules is given in [6].

D. Complex Fuzzy Classes

The FSM contains several complex fuzzy classes permitting to implement the semantics of real-world in terms of generalization, specialization, aggregation, grouping and composition relationships, which are commonly used in semantic modeling. A detailed description of these constructs is given in [5].

E. Computing the Degree of Membership

The degree to which each decision rule determines the fuzzy class K is not the same. To ensure this, each decision rule j is associated with a non-negative weight w_j reflecting its importance in deciding whether or not an entity e is a member of a given fuzzy class K . The d.o.m of entity e in fuzzy class K is computed as follows [5]: $\mu_K(e) = \frac{\sum_{j=1}^n \rho_j(v) \cdot w_j}{\sum_{j=1}^n w_j}$, where n is the number of decision rules, $v \in D(attr)$ and $\rho_j : D(attr) \mapsto [0, 1]$ is the *partial membership function* associated with the j th decision rule; it maps the elements of $D(attr)$ into $[0, 1]$ ($attr$ is the attribute on which the decision rule is based). For semantic decision rules, v is a semantic phrase and the partial d.o.m $\rho_j(v)$ is supposed to be equal to 1 but the user may explicitly provide a value less than 1. This basic definition of the d.o.m is used to define the membership degrees of complex fuzzy classes (see [5] for details).

III. MAPPING FSM TO FUZZORM DATABASE MODEL

Let M be a model based on FSM. The objective of the mapping process is to create a FuzzORM database model T that captures all the semantics of M . The mapping process consists in a set of *mapping rules* to be applied on the attributes, decision rules, simple and complex classes, and semantic relationships.

A. Mapping of Attributes

Attributes in FSM can be crisp or fuzzy. The list of fuzzy data types supported by FSM are given in Table I (see also [7]). A crisp attribute is basically characterized by its name, description, domain and data type. There are other system attributes but only these ones are considered in this paper. In addition to its basic data type, a fuzzy attribute is characterized by a set of parameters permitting to generate its possibility distribution. The number of parameters needed to define fuzzy attributes is different from one data type to another.

TABLE I. FUZZY DATA TYPES.

Type	Name	Example
1	Single scalar	quality= average
2	Simple number	age=30
3	Set of possible scalar assignments	quality={bad,average,good}
4	Set of possible numeric assignments	age={20,21,22,23}
5	Fuzzy range	age=between 20 and 30
6	Approximate value	age=about 35
7	Interval	age \in [25, 35]
8	Less/More than value	age=less/more than 35
9	Poss. dist. over a numeric domain	age={0.5/20,1.0/21,0.7/22,0.3,23}
10-13	Linguistic label (four models)	age=young
14	An unknown value	age=unk
15	An undefined value	age=und
16	A Null value	age=null

Figure 2 provides the graphical representation of the possibility distribution of three examples of fuzzy attributes. The Fuzzy Range that handles ‘more or less’ information between

two numeric values requires four parameters (α, β, γ and λ). The Approximate Value is defined by three parameters (c, c^- and c^+). The Interval data type is defined through the limits of the range α and β .

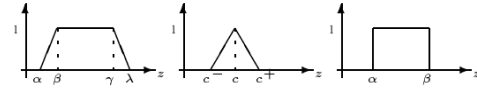


Figure 2. Graphical representation of some fuzzy data types.

In FuzzORM, the fuzzy attributes are mapped into composed and/or multi-valued attributes (supported by object relational database models) that store all the required parameters. This solution is formalized through several mapping rules. The following are two examples.

Mapping Rule 1. Let $attr$ be a crisp attribute in M . Attribute $attr$ is mapped to a new attribute with the same characteristics as in conventional databases.◊

Mapping Rule 2. Let $attr$ be a fuzzy attribute in M . The attribute $attr$ is mapped to a new attribute with the same characteristics plus an additional compound attribute *Parameters* with the following components: (i) *Value*, which is the value of the attribute as provided by the user; (ii) *DataType*, which is the fuzzy data type of the attribute provided by the user; and (iii) *ParametersList*, which is a multi-valued attribute indicating the list of parameters’ values needed to generate the possibility distribution of the fuzzy data type.◊

B. Mapping of Decision Rules

The characteristics of each decision rule should be mapped into the metadata level of FuzzORM. For attribute-based decision rules, we need to maintain the following information: (i) *RuleID* that stores the identifier of the decision rule; (ii) *RelationID* that indicates the name of the class to which the decision rule is associated; (iii) *DecisionRule*, which is a composite attribute defined as follows: (a) *AttrID* that references the attribute on which the decision rule is based, (b) *Operator* that contains a binary or a set operator, and (c) *RHO*, which is a crisp or fuzzy value from the attribute domain representing the Right-Hand Operand of the decision rule; and (iv) *Weight* that stores the weight of the decision rule as specified in the data model M .

The semantic decision rules are mapped similarly. However, in this case the attribute *DecisionRule* contains two components: (a) *Operator*, which is any semantic operator such as ‘IS-A’, ‘A-SET-OF’ or ‘A-PART’; and (b) *RHO*, which is a semantic phrase.

The mapping of decision rules from M to T is formalized as follows.

Mapping Rule 3. Let r_j be an attribute-based or semantic-based decision rule in the data model M . Then, the characteristics of decision rule r_j are mapped into the database model T as indicated above.◊

C. Transformation of Basic Fuzzy Classes

Each fuzzy class in the FSM model is mapped into a relation. The fuzzy attributes are mapped as explained in Section III-A. The crisp attributes are treated as in conventional databases. An additional non printable attribute, *DOM*, used

to store the global d.o.m is systematically added to the new relation. The decision rules associated with K are mapped as indicated in Section III-B.

The mapping of simple fuzzy classes is formalized as follows.

Mapping Rule 4. Let K be a fuzzy class in M with the attributes $attr1, \dots, attrp$. The mapping of K from M to T is as follows: (i) create a relation $R=(attr1', \dots, attrp')$ where $attr1', \dots, attrp'$ are the mapping of attributes $attr1, \dots, attrp$ using Mapping Rules 1 and 2; (ii) add an attribute DOM with real data type to R ; (iii) add the characterisers of each decision rule of K to T using Mapping Rule 3; and (iv) associate to the new relation R the triggers permitting to compute the d.o.m and to control the parameters of fuzzy attributes. ◊

The last operation will be discussed in Section IV.

D. Transformation of Subclass/Superclass Relationships

A fuzzy subclass B of a fuzzy superclass A in FSM is mapped in FuzzORM into a fuzzy relation that inherits all attributes of the fuzzy relation issued from A . In addition to the attribute DOM , the relation B contains a new attribute, denoted by $DOM-A$, which is used to store the d.o.m of one entity from fuzzy subclass B in its fuzzy superclass A . A fuzzy subclass in FSM may be attribute-defined, roster-defined or set-operation-defined. An attribute-defined fuzzy subclass has one or several attribute values that are in accordance with some discriminative values that characterize perfectly its members. For instance, the fuzzy subclasses NOVA and SUPERNOVA in Figure 1 are specializations of the fuzzy class STAR based on the attribute *type-of-star*. A roster-defined fuzzy subclass is simply defined by an explicit enumeration of its members. A set-operation-defined fuzzy subclass may be defined as the set-difference or the set-intersection of two or more fuzzy classes.

The following mapping rule formalizes the mapping of attribute-defined subclass/superclass relationships.

Mapping Rule 5. Let B be a fuzzy subclass of a fuzzy superclass A . The mapping of subclass/superclass relationship from M to T is as follows: (i) fuzzy classes A and B are transformed according to Mapping Rule 4; (ii) a new attribute, denoted by $DOM-A$ used to store the d.o.m of one entity from fuzzy subclass B in its fuzzy superclass A is added to the relation mapped from B ; and (iii) add to the database model T the definition parameters of subclass/superclass relationship (i.e., list of attributes used to categorize the elements of fuzzy class B). ◊

Similar mapping rules have been defined to transform roster-defined and set-operation-defined fuzzy subclasses. The main changes concerns the last operation. For roster-defined subclasses, the parameters are simply the list of the members of the fuzzy class B as specified by the user. The mapping of set-operation-defined fuzzy subclasses is more complicated since the fuzzy subclass has at least two fuzzy superclasses. Hence, the mapping rule above should be applied to each of these fuzzy superclasses. We need also to maintain the set operator used to define the subclass/superclass relationship.

E. Transformation of Interaction Relationships

An interaction relationship relates members of one fuzzy class to other members of one or many fuzzy classes. Let B_1, \dots, B_n be n fuzzy classes related by an n -ary interaction

relationship. Each participant fuzzy class has $n - 1$ attributes for relating each of its members to each of the other members. When a participant fuzzy class B_i is mapped into a relation in the database level, a composite attribute *InteractionList* is added to it. The *InteractionList* contains as many component attributes as the number of participant fuzzy classes. These component attributes are used to indicate the list of the related members from the other fuzzy classes.

On the other hand, when an interaction relationship requires the creation of new attributes, a new fuzzy interaction class is generated. In addition to its own attributes (that are specified in the interaction relationship), the new interaction class should contain the following attributes: (i) the key attributes of the classes participating in the interaction relationship; and a *DOM* attribute as for basic classes.

The mapping of interaction relationships is formalized as follows.

Mapping Rule 6. Let B_1, \dots, B_n be n fuzzy classes related by an n -ary interaction relationship in M . The interaction relationship is mapped as follows: (i) fuzzy classes B_1, \dots, B_n are mapped into relations R_1, \dots, R_n according to Mapping Rule 4; (ii) add to each relation R_i issued from fuzzy class B_i a composite attribute *InteractionList* for relating each of its members to the members of the other classes; and (iii) if the interaction relationship requires the creation of new attributes, then a new interaction relation is created as indicated above. ◊

A fuzzy class may participate in several relationships. In this case, the mapping of this fuzzy class requires as many composite attributes *InteractionList* as the number of interaction relationships.

F. Transformation of Fuzzy Complex Classes

As mentioned above, FSM supports several complex fuzzy classes (composite, aggregate or grouping classes). These classes are first mapped according to Mapping Rule 4. Then, we need to add to the metadata repository the characteristics of the semantic relationships. For instance, the composition relationships are characterized by the following information: (i) *RelationName1* that stores the name of the first fuzzy class; (ii) *RelationName2* that stores the name of the second fuzzy class; (iii) *DefinitionType* that indicates the way the composition relationship is defined (attribute-defined or enumerated); (iv) *Parameters*, which is a multi-valued attribute that stores the parameters associated with *DefinitionType* attribute (list of attributes or members); and (v) *DOM* that stores the d.o.m of the fuzzy relation named *RelationName1* in the fuzzy relation named *RelationName2*.

A collection of other mapping rules have been defined to transform the different fuzzy complex classes but they are not presented here. However, some examples are given in Section IV-B.

IV. IMPLEMENTATION

A prototype named Fuzzy Interface Module (FIM) supporting the different mapping rules is under development over the ORBMS PostgreSQL. The mapping of a FSM model into FuzzORM concerns three different levels: (i) *database system level*, which is associated with extended data manipulation languages devoted to handle different fuzzy operations that the database system should support; (ii) *database level*, which is

concerned with the way the imperfect information is internally stored. This concerns both attribute values and extensional definition of different relations; and (iii) *metadata level*, which concerns the intensional definition of relations. The metadata level is normally managed by the host database system. However, the full implementation of the FuzzORM model requires some additional metadata, which are not supported by current database systems. In what follows, we focus on the second and third levels only.

A. Database Level

The mapping rules detailed in Section III are traduced into a set of PL/SQL routines. The main routine takes a FSM model as a text file and generates a new FuzzORM database using the different mapping rules. The input text file is scanned three times. In the first scan, FIM identifies and implements the fuzzy domains, the decision rules, the metadata about the attributes and the metadata about the semantic relationships in the FSM model. In the second scan, FIM maps all the fuzzy classes (simple or complex) and implements them without considering the semantic relationships between the classes. In the third scan, FIM adds the semantic relationships.

The fuzzy classes are mapped into relations as detailed in Section III. The mapping of the FSM model in Figure 1 leads to the FuzzORM database given in Figure 3.

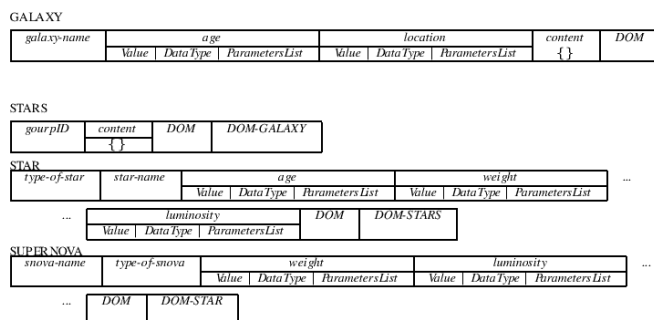


Figure 3. The FuzzORM Database.

As we can see, each fuzzy class is mapped into a relation and each fuzzy attribute into an attribute with composite type. The multi-valued attribute *content* is used in some complex fuzzy classes to maintain the list of members of these classes. This attribute is empty during the creation of the relations.

Each fuzzy relation is associated with several triggers to control the validity of the introduced data (i.e., the parameters of fuzzy attributes). Figure 4 provides a trigger example.

```
CREATE FUNCTION GalaxyLoc() RETURNS trigger AS $GalaxyLoc$
DECLARE
alpha numeric; beta numeric; gamma numeric; delta numeric;
BEGIN
IF NEW.location IS NULL THEN
RAISE EXCEPTION 'location cannot be null';
END IF;
IF NOT ((alpha<=beta) AND (beta<=gamma) AND (gamma<=lambda))
THEN
RAISE EXCEPTION 'Please provide valid parameters';
END IF;
END;
$GalaxyLoc$ LANGUAGE plpgsql;
CREATE TRIGGER GalaxyLoc BEFORE INSERT OR UPDATE ON Galaxy
FOR EACH ROW EXECUTE PROCEDURE GalaxyLoc();
```

Figure 4. A trigger example.

The code in Figure 4 represents a simple version of the trigger associated with the relation issued from fuzzy class GALAXY.

B. Metadata Level

The metadata contains several meta-relations for storing the different parameters and elements of FuzzORM such as fuzzy attribute characteristics, decision rules and semantic relationships. We define a meta-relation named FUZ-ATTRIBUTES to store the information about fuzzy attributes: their basic type, fuzzy types and their parameters. Here, we use a single column to define the parameters of fuzzy attributes, independently of the number of used parameters (1, 2, 3 or 4). In fact, the ORDBMS allows the use of multi-valued attributes permitting to maintain several values for a given attribute. This is not allowed in relational database systems where we should use 4 different columns to define the parameters of different fuzzy data types. This may cause many NULL values (when the number of parameters is less than 4). An extract from FUZ-ATTRIBUTES is given in Table II.

TABLE II. META-RELATION FUZ-ATTRIBUTES.

AttrID	AttrName	FuzzyDataType	BasicDataType	Parameters
1	luminosity	Fuzzy Range	Real	{0.05,0.2,1.0,1.3}
2	weight	Interval	Real	{1.2,1.5}
3	location	Linguistic Label	Text	{7.5,10}

The meta-relation LABELS is used to handle the characteristics of linguistic labels. An extract from the meta-relation LABELS is given in Table III.

TABLE III. META-RELATION LABELS.

LabelID	AttrID	Label	Parameters
1	7	very young	{0,1.8}
2	7	young	{1.5,5.0}
3	7	old	{4.2,11.3}
4	7	very old	{11,15}

The labels shown in this meta-relation are relative to the fuzzy attribute *age* (which is defined as a set of Gaussian linguistic labels) associated with fuzzy class STAR.

We also define other meta-relations for handling proximity relations, possibility distributions, domains of attributes, etc.

The metadata level contains also all the information required to define the decision rules associated with different fuzzy classes. They are stored in two meta-relations: A-DECISION-RULES and S-DECISION-RULES. The first one is devoted to store the definition of attribute-based decision rules and the second one is used to store the definition of semantic decision rules. The extensional definition of the meta-relations A-DECISION-RULES and S-DECISION-RULES for our example are given in Table IV and V, respectively.

There are also a set of meta-relations to handle the semantic relationships of subclass/superclass, composition, aggregation and grouping. The first meta-relation is SUB-SUPER-COMP that is devoted to store information concerning fuzzy subclass/superclass and composition relationships. The second meta-relation GROUPING is devoted to store information concerning grouping relationships. The meta-relations SUB-SUPER-COMP and GROUPING are given in Table VI and Table VII, respectively.

TABLE IV. META-RELATION A-DECISION-RULES.

RuleID	RelationID	RuleDefinition AttrID Operator RHO	Weight
1	STAR	{luminosity, ≥, 0.5L _s }	0.80
2	STAR	{weight, ≥, 0.05W _s }	0.30
3	SUPERNOVA	{luminosity, ≥, high}	0.60
4	SUPERNOVA	{weight, ≥, 1W _s }	0.50
5	NOVA	{luminosity, ≥, 0.5L _s }	0.80
5	NOVA	{weight, ≥, 0.05W _s }	0.30

TABLE V. META-RELATION S-DECISION-RULES.

RuleID	RelationID	RuleDefinition OperatorID RHO	Weight
1	GALAXY	{IS-A, Galaxy}	1.0

TABLE VI. META-RELATION SUB-SUPER-COMP.

Relation1Name	Relation2Name	RelationshipType	DefinitionType	Parameters	DOM
GALAXY	PLANETS	Aggregation	Enumerated	{}	1.0
GALAXY	STARS	Aggregation	Enumerated	{}	1.0
GALAXY	COMETS	Aggregation	Enumerated	{}	1.0
STAR	SUPERNOVA	Subclass/ Superclass	Attribute	{star-type}	0.9
STAR	NOVA	Subclass/ Superclass	Attribute	{star-type}	1.0

TABLE VII. META-RELATION GROUPING.

RelationName	DefinitionType	Parameters	DOM
STARS	Enumerated	{}	1.0

The multi-valued attribute *Parameters* is used to maintain the list of members of complex fuzzy classes defined by enumeration. The list of members will be specified by the user progressively during the exploitation of the database.

C. Fuzzy Operations

To compute the membership degrees and for query processing, we need to extend the binary and the set operators that may be used in the definition of decision rules. An attribute-based decision rule is associated with a condition of the form:

$$\langle attr \rangle \langle op \rangle \langle v \rangle$$

The operator *op* may be a binary operator (i.e., =, ≈, ≤, <, ≥, >) or a set operator (i.e., ∈, ⊆, ⊂, ⊇, ⊃). All these operators may be associated with the negation operator, denoted ‘¬’ below. In conventional logic, the response to a binary comparison is a two-valued one and may be true (1) or false (0). Within fuzzy logic, the result of a comparison may take any value in the range [0,1]. Thus, the two-valued logic is simply a special case of fuzzy logic that is restricted to the two extreme values (0 and 1) of the range [0,1].

Based on the work of [8], we propose an extension of all the operators mentioned above to support fuzzy logic. For instance, the fuzzy operator ‘≈’, which gives the degree in which two fuzzy numbers (approximate values in Table 1) are approximately equal is defined as follows:

$$\mu_{\approx}(\tilde{x}, \tilde{y}) = \begin{cases} 0, & |\tilde{x} - \tilde{y}| > \text{margin}; \\ 1 - \frac{|\tilde{x} - \tilde{y}|}{\text{margin}}, & |\tilde{x} - \tilde{y}| \leq \text{margin}. \end{cases}$$

Here, we suppose that the parameters c^+ and c^- of an approximate value (see Figure 2) are the same and equal to *margin*. The fuzzy ‘¬ ≈’ operator is computed as the complement of ‘≈’ operator, i.e., $\mu_{\neg \approx} = 1 - \mu_{\approx}(\tilde{x}, \tilde{y})$.

D. DOM Computing Routines

FIM proposes a set of routines for computing membership degrees. These routines are defined as triggers and associated with the corresponding relations. The basic routine named DOM is used to compute global and partial membership degrees associated to the fuzzy instances of classes (tuples in database). This routine permits to compute $\mu_K(\cdot)$ (see section II-E) and is associated to the INSERT and UPDATE triggers. The principle of DOM routine is given in Figure 5.

Algorithm 1: DOM

```

Input : R, // relation.
         t, // tuple.
Output: real ∈ [0, 1] // d.o.m.
P ← {set of decision rules for R};
W ← {decision rules weights};
wdom ← 0;
wsum ← 0;
for (each r ∈ P) do
  wdom ← wdom + W[r] * RHO(r, t);
  wsum ← wsum + wdom;
if (wsum = 0) then
  return 0;
return wdom / wsum;
    
```

Figure 5. The d.o.m computing algorithm.

The DOM routine uses the RHO function, which permits to calculate the partial membership degrees (see Section II-E).

V. DISCUSSION AND RELATED WORK

In this section, we discuss some implementation issues and some related work.

A. Discussion

The first implementation issue to discuss concerns the mapping and storing of fuzzy attributes’ parameters. There are several solutions to map fuzzy attributes. We can, for example, use one common meta-relation with four attributes devoted to store the different parameters. In that time, we may have ‘null’ values any time the number of parameters is less than four. Another solution is to group data types along the number of required parameters. After that, four relations are needed for data types with one, two, three or four parameters, respectively. An ameliorated version of this solution is adopted in [8] where a common meta-relation is defined with a specific attribute serves as a pointer to two other meta-relations. One drawback of the solutions cited above is that anytime we need to add a new linguistic data type or to change the adopted linguistic data type, we may have to update the meta-relations structure.

The straightforward solution proposed in this paper does not depend on the parameters number because it uses multi-valued attributes allowing the storage of more than one single value. That is, all the needed parameters of linguistic data type may be defined using only one attribute. This solution has several advantages: (i) is supported by object relational database models; (ii) reduces the presence of ‘null’ values; and (iii) the atomic attributes of a composite attribute remain accessible individually as with non-composite attributes.

The second issue concerns the use of the attribute *Parameters* both at the intensional and extensional levels. This allows users to insert values of different data types, which may have different number of parameters. For instance, the formal definition of the attribute may be a trapezoidal-based

possibility distribution with four parameters but the user may introduce a crisp value (with no parameter at all), an interval (with two parameters) or an approximate value (with three parameters). In all cases, the different data types defined at the extensional level should be consistent with the formal definition of the attribute at the intensional level.

The third issue is related to the computing of the d.o.m. In FSM model, we need to compute the partial and global membership degrees associated with the class instances. At the database level, we need to compute the partial and global membership degrees of any tuple. Three solutions may be adopted: (i) store the partial membership degrees and compute global membership degree on the fly; (ii) store the global membership degree and compute partial membership degrees on the fly; and (iii) store both partial and global membership degrees. The first solution is expensive in access time because we need to compute the global membership degree frequently. The third solution may ameliorate the access time substantially but this needs, however, a much more storage space. In this paper, we use the second solution as it is less expensive and ameliorate access time substantially. Equally, the second approach where we compute global membership degree on preprocessing allows us to use global membership degrees as a filter to eliminate quickly false alarms in the querying process. This may reduce execution time, especially for large databases.

B. Related Work

Fuzzy information has been extensively investigated in the context of relational database model [1][2]. There are also several semantic [5][1][9][2][10][11][12] and object-oriented [13][14] database models where fuzziness is introduced with one or several levels. Due to the lack of effective implementation mechanisms, most of fuzzy semantic data models have been mapped and implemented through relational database models. For instance, the well-known Entity-Relationship (ER) data model is extended to support fuzziness in [1]. The paper includes also a fuzzy entity-relationship methodology for the design and development of fuzzy relational databases. This methodology was used for mapping the fuzzy ER to a relational one. In [15], the Is-a relationships, Functional relationships, complex Objects (IFO) model was extended to the Extended IFO (ExIFO) to represent uncertainty as well as precise information. The authors provide also an algorithm for mapping the schema of the ExIFO model to an extended NF² database model. In [2], the IFO data model is extended to support fuzziness. The obtained model, denoted IF₂O, is then mapped to a relational fuzzy database schema.

There are also some proposals for mapping semantic data models into object-oriented database models, which permit to support several concepts of semantic modeling. For instance, in [4] the authors extend the IFO model for handling ill-defined values including values with semantic representation, values with semantic representation and conjunctive meaning, values with semantic representation and disjunctive meaning. The paper includes also a mapping of the obtained fuzzy IFO model to a fuzzy object-oriented database model. The proposal of [3] presents an extension of the Extended Entity-Relationship (EER) model to deal with fuzzy information. The paper provides also a formal design methodology for fuzzy object-oriented databases from the fuzzy EER model.

VI. CONCLUSION

This paper provides a formal approach for mapping the FSM to a FuzzORM database model. Compared to existing proposals, our mapping approach permits to gather advantages of both relational and object-oriented database models. Several points related to this proposal need to be addressed: (i) the enrichment of the semantics of FSM through the addition of different constraints associated with attributes, entities and classes; and then their incorporation into FuzzORM; (ii) the definition and implementation of a data definition language adapted to FuzzORM databases; (iii) the implementation of the conceptual query language introduced in [5]; and (iv) the enrichment of FIM with capabilities related to the definition of imperative aspects of the database such as security, integrity and access levels.

REFERENCES

- [1] N. Chaudhry, J. Moyne, and E. Rundensteiner, "Extended database design methodology for uncertain data management," *Information sciences*, vol. 121, no. 1, 1999, pp. 83–112.
- [2] Z. Ma, "A conceptual design methodology for fuzzy relational databases," *Journal of Database Management*, vol. 16, no. 2, 2005, pp. 66–83.
- [3] Z. Ma, W. Zhang, W. Ma, and G. Chen, "Conceptual design of fuzzy object-oriented databases using extended entity-relationship model," *International Journal of Intelligent Systems*, vol. 16, no. 6, 2001, pp. 697–711.
- [4] M. Vila, J. Cubero, J. Medina, and O. Pons, "A conceptual approach for dealing with imprecision and uncertainty in object-based data models," *International Journal of Intelligent Systems*, vol. 11, no. 10, 1996, pp. 791–806.
- [5] R. Bouaziz, S. Chakhar, V. Mousseau, S. Ram, and A. Telmoudi, "Database design and querying within the fuzzy semantic model," *Information Sciences*, vol. 177, no. 21, 2007, pp. 4598–4620.
- [6] L. Ellouze, R. Bouaziz, and S. Chakhar, "Extending fuzzy semantic model by advanced decision rules," in *Annual Meeting of the North American Fuzzy Information Processing Society, 2009. NAFIPS 2009, June 2009*, pp. 1–6.
- [7] A. Bahri, S. Chakhar, Y. Naja, and R. Bouaziz, "Implementing imperfect information in fuzzy databases," in *Proceedings of The International Symposium on Computational Intelligence and Intelligent Informatics, October 14-16 2005*, pp. 1–8.
- [8] J. Medina, M. Vila, J. Cubero, and O. Pons, "Towards the implementation of a generalized fuzzy relational database model," *Fuzzy Sets and Systems*, vol. 75, no. 3, 1995, pp. 273–289.
- [9] G. Chen and E. Kerre, "Extending ER/EER concepts towards fuzzy conceptual data modeling," in *Fuzzy Systems Proceedings, 1998. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference on*, vol. 2, May 1998, pp. 1320–1325.
- [10] Z. Ma, F. Zhang, L. Yan, and J. Cheng, "Fuzzy data models and formal descriptions," *Studies in Fuzziness and Soft Computing*, vol. 306, 2014, pp. 33–60.
- [11] L. Yan and Z. Ma, "Incorporating fuzzy information into the formal mapping from web data model to extended entity-relationship model," *Integrated Computer-Aided Engineering*, vol. 19, no. 4, 2012, pp. 313–330.
- [12] F. Zhang, Z. Ma, and L. Yan, "Representation and reasoning of fuzzy ER models with description logic DLR," *Journal of Intelligent and Fuzzy Systems*, vol. 26, no. 2, 2014, pp. 611–623.
- [13] Z. Ma, W. Zhang, and W. Ma, "Extending object-oriented databases for fuzzy information modeling," *Information Systems*, vol. 29, no. 5, 2004, pp. 421–435.
- [14] C. Cuevas, N. Marin, O. Pons, and M. Vila, "Pg4DB: A fuzzy object-relational system," *Fuzzy Sets Systems*, vol. 159, no. 12, Jun. 2008, pp. 1500–1514.
- [15] A. Yazici, B. Buckles, and F. Petry, "Handling complex and uncertain information in the ExIFO and NF² data models," *IEEE Transactions on Fuzzy Systems*, vol. 7, no. 6, 1999, pp. 659–676.

Knowledge and Technology Transfer in the Center for Scientific and Technical Information of the Wrocław University of Technology

Anna Walek

Center for Scientific and Technical Information
Wrocław University of Technology
Wrocław, Poland
anna.walek@pwr.edu.pl

Katarzyna Kozłowska

Center for Scientific and Technical Information
Wrocław University of Technology
Wrocław, Poland
katarzyna.kozłowska@pwr.edu.pl

Abstract—Knowledge and Technology Transfer between a university and economic operators affects innovation and growth of competitiveness, as well as the development of a knowledge-based society. In the structures of Wrocław University of Technology, regarded as one of the best and the most innovative technical universities in Poland, a number of units responsible for a wide understanding cooperation with the economy have been established. To avoid dispersion of various activities, competences and information, Wrocław University of Technology has started to implement a technology transfer system. A Contact Point for Technology Transfer has been established in the Center for Scientific and Technical Information. It coordinates activities of all the organizational units of the University, as well as people who fulfill tasks concerning technology transfer, commercialization and cooperation with the economy. The paper discusses the activities of the Center for Scientific and Technical Information related to knowledge and information transfer, including guidance of commercializing research results, patent information and assistance in the purchase of innovative solutions, as well as intellectual property generated at Wrocław University of Technology.

Keywords—Wrocław University of Technology; technology transfer; knowledge transfer; commercialization.

I. INTRODUCTION

Knowledge transfer (KT) is a term used to include a very broad range of activities to support mutually beneficial collaborations between universities, businesses and the public sector.

It's all about the transfer and intellectual property (IP), expertise, learning and skills between academia and the non-academic community. For academics, KT can be a way of gaining new perspectives on possible directions and approaches to research. Discussion around KT often focuses on the formation of spin-out business, or the licensing of intellectual property, based on the outputs of university scientific and technology-related research. Although these are vitally important areas, KT actually encompasses a much broader range of activities and is not limited to science and technology disciplines [1].

The underlying assumption that there is a potential for increased collaboration between industry and universities is also underlined in much of current innovation literature. In particular, the Open Innovation [2] approach to developing business value is explicitly based on an assumption that Universities are a "vital source for accessing external ideas". Moreover Universities have been deemed to be "the great, largely unknown, and certainly underexploited, resource contributing to the creation of wealth and economic competitiveness" [3].

Nowadays, the term "knowledge transfer" is usually applied to transfers between universities and industry, or between experts and non-experts. The focus is frequently on "science management" [4].

Technology transfer, also called transfer of technology (TOT), is the process of transferring skills, knowledge, technologies, methods of manufacturing, samples of manufacturing and facilities among governments or universities and other institutions to ensure that scientific and technological developments are accessible to a wider range of users who can then further develop and exploit the technology into new products, processes, applications, materials or services [5].

According to the definition in the "Business Dictionary", technology transfer is "assignment of technological intellectual property, developed and generated in one place, to another through legal means such as technology licensing or franchising" [6].

The process typically includes: identifying new technologies; protecting technologies through patents and copyrights; forming the development and commercialization strategies, such as marketing and licensing to existing private sector companies or creating new startup companies based on the technology.

Section II of the paper presents the background of creating the knowledge transfer and transfer of technology at Polish universities, which is connected with current situation of technical universities. Sections III and IV describe Wrocław University of Technology and the Center for Scientific and Technical Information (CSTI) along with its' departments responsible for KT and TOT. Their tasks and the ways of implementing university system of cooperation

with the economy have been depicted. Section V presents detailed results of CSTI activity connected with the cooperation with the economy. This section presents work effects of particular teams and the specifics of inquiries and commissions directed to CSTI departments from the business and external entities. Last but not least, at the end of the paper, tables and graphs to illustrate the content of this paper have been presented. They depict commercialization scenarios, models and patterns of particular forms of cooperation with the economy, which have been developed and implemented at Wrocław University of Technology.

II. BACKGROUND

The importance of creating networks between business entities, public administration, non-governmental, scientific and research institutions is growing steadily. Such networks help merging ideas, exchanging information and establishing cooperation methods between the above. The innovation and entrepreneurship centers which have been developing in Poland since the early 1990s of the 20th century are gaining in importance in these processes. Currently, the entrepreneurship support infrastructure in Poland consists of different types of innovation and entrepreneurship centers: technology parks and incubators, business incubators, pre-incubators, technology transfer centers, training and advisory centers, loan funds, guarantee funds, seed funds etc. These institutions are generally intended to enhance human creativity, entrepreneurship and innovation leading to more effective use of the local growth factors. Since the beginning of system transformation, the number of innovation and entrepreneurship centers has been systematically growing. The process of developing the entrepreneurship supporting system is still running. New initiatives and new areas where the innovation and entrepreneurship centers operate are appearing. The changes observed and analysis of the experience acquired by “knowledge economy leaders” show an increasing role of support infrastructure in the process of the Polish economy innovation development. In the era of technological changes and dynamic expansion of the innovation to the services, organizations, marketing and social issues, the enterprises are looking for new solutions and this is where the innovation and entrepreneurship centers may have important contribution [7].

Nowadays, company competitiveness is based on components like knowledge, technological capabilities and skills. This led to a theoretical discussions on knowledge – based economies. In order to create this knowledge, which will eventually be transformed into new products and services, companies have internationally began to form increasing numbers of knowledge-based strategic alliances, thus creating a new form of competition. Nevertheless, the creation and transfer of knowledge and best practices through cooperation have proven to be quite difficult. Knowledge transfer is neither an easy nor a costless task. Unlike information, capabilities and knowledge simply cannot be bought in market. Instead, they have to be gradually built through intensive and systematic learning efforts. The issue of knowledge management and knowledge

transfer is all the more important for firms lacking, or late, in technological capabilities [8].

III. WROCLAW UNIVERSITY OF TECHNOLOGY

Wrocław University of Technology is one of the best technical universities in Poland. It was founded in 1946 in Wrocław, and organized by researchers from pre-war research centers in Lvov and Warsaw. Since the very beginning of its existence, it has been an important center of technical education. Today, over 34 000 students study here under the guidance of 2 000 academic teachers, at the 12 faculties, as well as in the 3 regional branches (Jelenia Góra, Legnica, Wałbrzych). It rates high in the annual rankings of Polish universities.

Wrocław University of Technology is situated in Lower Silesia – a dynamically developing region of Poland. Focusing on adopting its own offer to the market needs has become the regional strategy. Due to that, a synergy effect has been achieved with the development of segments strategic for the region. The University is strongly oriented toward cooperation with the economy and industry [11].

An excellent geographic location, teaching and research backup, and the developing infrastructure are the key assets of Lower Silesia, which have convinced international corporations to make investments there. Projects by such companies, as Volvo, Toyota, Volkswagen, Whirlpool, WABCO, Siemens, and LG Philips, 3M and Toshiba, have been implemented in the recent years in Lower Silesia.

Crucial for attracting investors to the region is the research and scientific potential of Wrocław academic center and the work on creating a knowledge-based economy. Wrocław University of Technology, as the only technical university in the region, has become a leader of active cooperation with the industry. Collaboration with the economy allows providing a comprehensive offer for companies looking for innovative solutions. Tens of long-term, many-sided collaboration agreements signed prove that this is the right approach to business partners. These agreements cover a wide range of activities, from providing training for the employees of Lower Silesian companies to joint research aiming at introducing new technologies and products, and increasing company competitiveness on Polish and international markets [12].

IV. CENTER FOR SCIENTIFIC AND TECHNICAL INFORMATION

On January 1st 2014, Wrocław University of Technology launched the Center for Scientific and Technical Information (CSTI). It is a unit serving the whole university performing scientific, research and service-oriented tasks. The Center is responsible for collecting and providing scientific and technical information for the needs of performing scientific research and supporting didactics, as well as coordinating cooperation with the economy and technology transfer. Within the structure of the Center, the Traditional and Electronic Libraries were established, providing the library-information services and creating also a digital library, knowledge repository and the data base for scientific

achievements. Besides the library resources, data bases and electronic periodicals, the Center makes available the patent and standardizing information, as well as the information on new technologies both, for the needs of scientific society and industry representatives [11].

Within the framework of the CSTI, units dedicated to cooperation of science with the economy, such as the Center for Science and Economy Cooperation (CSEC) operate. Its activity is focused on tasks supporting and initiating undertakings of all types in cooperation with representatives of business environment. The Center for Scientific and Technical Information runs the Contact Point for Technology Transfer. By creating a network of mutual relations with business and industry representatives, it identifies individual needs of enterprises in innovation, enabling that way development of solutions facilitating functioning of mechanisms of knowledge commercialization and widely understood cooperation of the University with businesses [13].

A. Center for Science and Economy Cooperation

The mission of the Center for Science and Economy Cooperation is to support the transfer of knowledge and information between the Wrocław University of Technology and external economic entities, with particular emphasis on the region of Lower Silesia, as well as to promote cooperation between academic and economic environment on regional and national level.

The activities of the CSEC focus on the tasks supporting and initiating all kinds of projects in cooperation between university units and business representatives. As a unit responsible for the coordination and organization of the cooperation process. The CSEC ensures efficient circulation of documents and deals with administrative formalities before signing cooperation agreements.

1) Department of Knowledge and Information Transfer

The Department of Knowledge and Information Transfer, which has been separated from the structure of the CSEC, undertakes activities, which aim at presenting University research offer. Moreover, these activities enhance the processes of efficient acquisition and transfer of know-how, allowing more effective use of knowledge by the economic environment.

The Department of Knowledge and Information Transfer realizes part of a strategic plan for the development of the University, involving creation of a new organizational model and entrance into qualitatively new relationship with business, thus increasing the possibilities for commercial exploitation of knowledge resources. The activities aim at creating optimal conditions and solutions for the use of the scientific and technical potential of the University in specific areas of the economy.

In the effort to ensure simplification and adjustment of the system regulation, as well as by direct consultancy, coordination, processing of orders and agreements resulting from the cooperation with third parties, the staff of the Department of Knowledge and Information Transfer counteract procedural constraints. Not only do they render

complex services, but they also complete all formalities on company behalf.

Taking into account the trend of development of the world economy based on the information society where innovation is a key factor, the staff of the Center for Science and Economy Cooperation have worked out a software which integrates databases identifying scientific potential of the University, including the access to expertise, cutting edge technologies, and research infrastructure. It is not only a perfect tool for promotion of intellectual output of the Wrocław University of Technology, but it also plays a role of a forum for exchange of modern technical and scientific ideas.

The collected data have been published on the website of the Center for Science and Economy Cooperation (<http://ofertadlagospodarki.pwr.edu.pl/en/o-nas/>) and include a variety of subject areas: research offer; laboratories (summarizes the list of research laboratories at the disposal of Wrocław University of Technology); key equipment; experts (the aim of creating this database was to provide both enterprises and scientists with a wide range of specialists, thus allowing efficient search for the right information, as well as contact); projects (on this site, users can publish new topics and attach suggestions for joint undertakings, resulting in starting new projects); inventions (it is a great tool for promoting research achievements of scientists among potential contractors interested in the implementation of products); standardization; knowledge repository (database developed in the Electronic Library. It includes scientific publications and papers, documentation of research data and other documents and resources being a result of scientific research and development works); product catalogue; innovative solutions (the database consists of technological solutions with full technical documentation, which can be immediately implemented in manufacturing companies); research centers.

Databases are regularly supplied with information verified by persons specialized in different disciplines. Databases do not only document scientific achievements of the University employees, but they also provide a direct exchange of information, enable its recording and processing, and in particular allow multivariate data mining.

The application mentioned above, accessible on the Center website, is extremely important from the point of view of the support for the commercialization process of the research results. One should take into account, that the scientific qualities of projects contribute often to the development of innovative products.

2) Department of Intellectual Property and Patent Information

The Department of Intellectual Property and Patent Information, within the CSEC, deals with problems of industrial property. It carries out activities for academic community of Wrocław University of Technology. Patent attorneys employed in the Department act as plenipotentiaries of the University in the proceedings before the Polish Patent Office, in cases related to obtaining and maintaining exclusive rights to solutions, which were

developed or co-developed by the University. Moreover, the Department comprises the Regional Standardization Information Center, which is a part of European PATLIB network, and runs consultancy activities concerning problems of intellectual property, its commercialization and cooperation with the regional economic environment. The Department staff provide patent information, help in use of patent databases and render legal advice not only to the staff of the Wrocław University of Technology but also to private individuals, thereby maintaining strong links between the University and economic entities.

In addition to providing professional service of patent attorneys and patent information, cooperation with industry reveals itself primarily in preparation of commercialization and substantive care throughout its course. An important task of the Department of Intellectual Property and Patent Information is concluding license agreements and contracts for the commonality of rights to industrial property objects and consulting draft agreements in terms of intellectual property with other departments, in particular the Department of Knowledge and Information Transfer. Implementations bring tangible and mutual benefits, contribute to the competitiveness of enterprises, increase their profits and favor economic growth. They are of use also to inventors. Thanks to attractive conditions of compensations, the Terms of Use of Intellectual Property, adopted at the University, motivate authors of inventions to support the commercialization process and search for potential partners.

B. Technology Transfer Contact Point

One of the first stages of implementing the System of Technology Transfer at Wrocław University of Technology was launching the Technology Transfer Contact Point (TTCP) within the structures of the Center for Scientific and Technical Information.

The aim of the Contact Point is to distribute powers and coordinate activities and cooperation of organizational units of the University in the commercialization process of intellectual property rights. In particular, the Contact Point excels in:

- identifying and monitoring research projects with high potential of commercialization of end products;
- supporting innovative ideas and technical and technological solutions in the process of gaining business partners for their development and applying for financing the implementation works;
- informing and consulting administrative and legal issues regarding opportunities and procedures of technology transfer;
- informing and consulting possibilities of obtaining financial support for the technology transfer process;
- rendering administrative and legal services related to starting and transferring intellectual property rights to new companies.

As a unit responsible for coordination of the actions mentioned above, the Contact Point divides tasks and responsibilities between organizational units of the University, and initiates and supports development and

implementation process of procedures and paths of commercialization, intellectual property management regulations, foundation of spin-off companies, internal regulations.

Launching the Contact Point will significantly improve and simplify procedures related to initializing and running commercialization actions at the University through coordination and implementation of the “one stop shop”.

The Contact Point for Technology Transfer should be perceived by scientists and entrepreneurs as the basic path of contact with the University on issues of technology transfer, not excluding previously used forms of communication - through Wrocław Center of Technology Transfer, departments or other organizational units. In addition, the Contact Point will be equipped with an electronic system of recording and monitoring orders of commercialization. It will allow to track the current status of ongoing actions for technology transfer at the University level.

The mission of the Technology Transfer Contact Point is coordinating activities of University departments as far as technology transfer is concerned. The Technology Transfer Contact Point is also responsible for overall registration of commercialization processes. Moreover, following the ‘one stop shop’ concept, it helps external entities contacting University units. Last but not least, it supports academic staff in establishing cooperation with business units in the area of technology transfer and through disseminating information about the development of new technical solutions. Each TTCP user will eventually save their time, as the TTCP staff will support them in fulfilling procedures and formalities connected with the commercialization process.

V. KNOWLEDGE AND TECHNOLOGY TRANSFER – PRACTICE

The main achievement of the Contact Point is the development and implementation of commercialization models and scenarios depicted below.

There are two scenarios of technology transfer in WrUT (Fig. 1).

Scenario 1 represents direct commercialization consisting in a direct sale to a company:

- The University has 30 days to make a decision whether they are interested in a given commercialization. If the decision is positive, the author gets informed and he/she is then to decide whether he/she wants to sign the contract with Wrocław University of Technology on proposed conditions. If he/she agrees, an agreement with the details of commercialization conditions and remuneration is concluded.
- The University renounces the commercialization or has not kept the 30-day-long deadline. In such case, the University transfers the offer to the author for maximum 10% remuneration. The author is then to decide whether he/she wants to commercialize individually. If yes, the author concludes an agreement transferring his/her economic rights with

the University and the commercialization is performed by the author.

Scenario 2 represents commercialization by means of a special purpose company:

In-kind contribution to a special purpose company (Fig.2):

- Deciding to commercialize a product and concluding a commercialization condition agreement with the author
- Determining copyrights distribution (the University covers the costs)
- Finding product recipient and makes product valuation (by the University or an external entity)
- Introducing products as University assets
- Obtaining a consent of the University for in-kind contribution
- Preparing and concluding an in-kind contribution contract
- Issuing and invoice and VAT settlement

First model of commercialization process described below (Option 2) by means of a special purpose company, depicting commercialization income path (Fig. 3):

- Sales revenues and company profits go to a special purpose company.
- University receives a revenue refund for technology transfer to the special purpose company.
- University pays remuneration to the author.

Second model of commercialization process (Option 1 – direct) depicting the path of commercialization incomes (Fig. 4):

- Economic entities transfer the sales and licensing income to the University
- The University pays remuneration to the author

The following tables present effects of the cooperation of the Center for Science and Economy Cooperation and the Contact Point with economic environment in 2014.

TABLE I. THE NUMBER OF DIFFERENT TYPES OF AGREEMENTS ELABORATED AND PROCEEDED IN THE CENTER FOR SCIENCE AND ECONOMY COOPERATION AND IN THE CONTACT POINT

No.	Elaborated agreements	Number
1.	Framework cooperation agreements	27
2.	Consortium agreements	10
3.	Research and development work agreement s	11
4.	Agreements on cooperation in framework of a project	11
5.	Financing agreements	4
6.	Non-disclosure agreements	4
7.	Agreements on a series of lectures	2
8.	Agreements on transfer of materials	2
9.	Donation agreements	1
10.	License agreements	1

No.	Elaborated agreements	Number
11.	Letters of intent	7
12.	Memorandums of understandings	3
TOTAL		83

The next table presents different requests addressed to the Center for Science and Economy Cooperation and to the Contact Point in 2014.

TABLE II. REQUESTS IN 2014

No.	Letter of inquiry	Number
1.	Requests for performing tests/analyses	83
2.	Requests for cooperation	11
3.	Requests for an expertise	4
4.	Requests for preparing a review of innovation	8
5.	Requests for preparing a review for the court	4
6.	Other	13
TOTAL		123

The CSEC is responsible also for registration of the research and development work agreements, which in fact are research work orders commissioned and paid by different companies and external institutions. In 2014, our Center registered 252 of such agreements.

The results of the activities undertaken by the Department of Intellectual Property and Patent Information for 2014 are shown in the following tables.

TABLE III. THE NUMBER OF CASES HANDLED BEFORE THE PATENT OFFICE ON BEHALF OF THE UNIVERSITY

No.	Application type	Number
1.	Prepared applications for innovative solutions to the Polish Patent Office or European Patent Office or Office for Harmonization in the Internal Market	169
2.	Prepared trademark applications	7
3.	Total number of prepared applications	176
4.	Applications to the Patent Office in defense of inventions (responses to allegations or other provisions)	178
5.	Patents and other exclusive rights granted to the University	130

TABLE IV. THE NUMBER OF AGREEMENTS WITH EXTERNAL ENTITIES IN THE FIELD OF INTELLECTUAL PROPERTY

No.	Type of technology transfer	Number
1.	Agreements on community of law	36
2.	License agreements and cession of rights agreements	6

The Department of Intellectual Property and Patent Information supports the University also in the scope of

supervising the agreements. In 2014, the Department gave opinions on or negotiated 439 agreements.

CONCLUSION

With the move of advanced economies from a resource-based to a knowledge-based production, many national governments have increasingly recognized “knowledge” and “innovation” as significant driving forces of economic growth, social development, and job creation. In this context the promotion of “knowledge transfer” has increasingly become a subject of public and economic policy. Within the framework of knowledge and technology transfer (KTT), the university aims to use new research findings in cooperation with industry – as quickly and efficiently as possible – in an attempt to produce new products and services and to positively shape social development.

Wrocław University of Technology takes steps towards developing a system of cooperation with the economy. The implementation of a “one stop shop” concept is to support and facilitate contacts between the representatives of the scientific community and the economy. Last but not least, the goal of the system is to facilitate the access to information on new technologies, as well as the cooperation of business community with the university.

REFERENCES

- [1] What is knowledge transfer? University of Cambridge. Research <http://www.cam.ac.uk/research/news/what-is-knowledge-transfer> [retrieved: 10 October, 2014].
- [2] L. Argote and P. Ingram, "Knowledge transfer: A Basis for Competitive Advantage in Firms", *Organizational Behavior and Human Decision Processes*, 82 (1), pp. 150–169, 2000. doi:10.1006/obhd.2000.2893
- [3] G. Holland, "Foreword", in H. Gray, "Universities and the creation of wealth", *The Society for Research into Higher Education and Open University Press*.
- [4] V. Lipphardt and D. Ludwig, "Knowledge Transfer and Science Transfer", in: *European History Online (EGO)*, published by the Institute of European History (IEG), Mainz 2011-12-12. URL: <http://www.ieg-ego.eu/lipphardtvludwigd-2011-en> URN: urn:nbn:de:0159-2011121229 <http://ieg-ego.eu/en/threads/theories-and-methods/knowledgetransfer/veronika-lipphardt-david-ludwig-knowledge-transfer-and-science-transfer>.
- [5] R. Grosse, "International Technology Transfer in Services". *Journal of International Business Studies* 27, pp. 782, 1996.
- [6] Business Dictionary, <http://www.businessdictionary.com/definition/technology-transfer.html>
- [7] Innovation and entrepreneurship centers in Poland. Report 2012, pp. 5, http://www.pi.gov.pl/PARPFfiles/file/OIB/IOB_Raporty_po_angielsku/2012_BSI_in_Poland_Report.pdf [retrieved: 14 October, 2014].
- [8] G. Tselekidis and A. Rafailidis, "Knowledge Transfer Management: studying the actual process", *The Fifth European Conference on Organizational Knowledge, Learning, and Capabilities 2-3 April, 2004, University of Innsbruck, Austria* http://www2.warwick.ac.uk/fac/soc/wbs/conf/olkc/archive/oklc5/papers/i-4_tselekidis.pdf
- [9] Perspectives University Ranking, http://www.perspektywy.pl/portal/index.php?option=com_content&view=article&id=724:uczelnie-akademickie&catid=93&Itemid=230 [retrieved: 11 October, 2014].
- [10] About the University. Wrocław University of Technology, http://www.portal.pwr.wroc.pl/info_podstawowe.242.dhtml [retrieved: 1 October, 2014].
- [11] A. Walek, "Center for Scientific and Technical Information - Library Services for Business and Science at Wrocław University of Technology. eKNOW 2014 : the Sixth International Conference on Information, Process, and Knowledge Management, March 23-27, 2014, Barcelona, Spain", [Wilmington, DE, USA] : IARIA, pp. 93-97, 2014.
- [12] Research. Wrocław University of Technology, <http://www.portal.pwr.wroc.pl/346196,242.dhtml>, [retrieved: 1 October, 2014].
- [13] „Development of a model of cooperation Science of Economics in Wrocław University of Technology”, Wrocław: Politechnika Wroclawska, 2014, unpublished.
- [14] Knowledge transfer. Wikipedia the free encyclopedia, http://en.wikipedia.org/wiki/Knowledge_transfer [retrieved: 5 October, 2014].

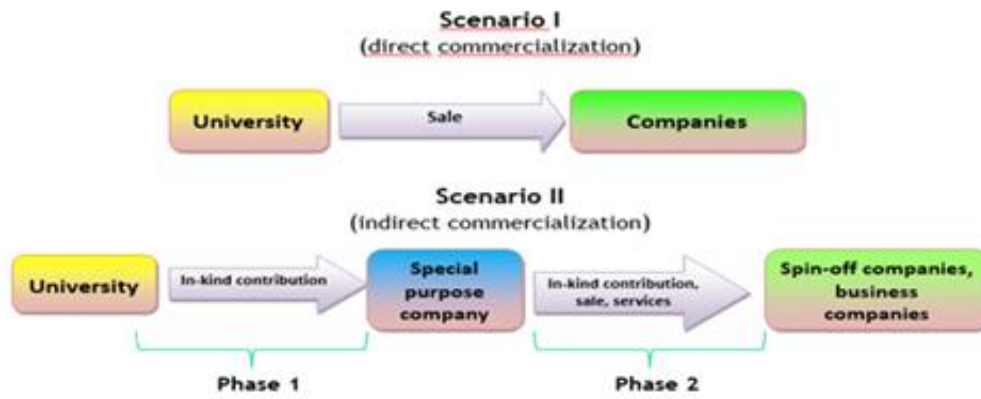


Figure 1. Technology Transfer in WUT – scenarios.

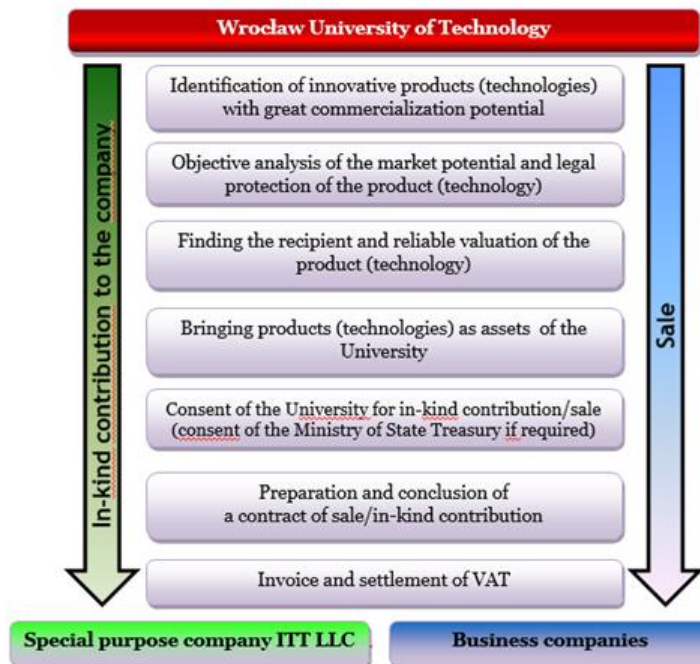


Figure 2. Model of commercialization process

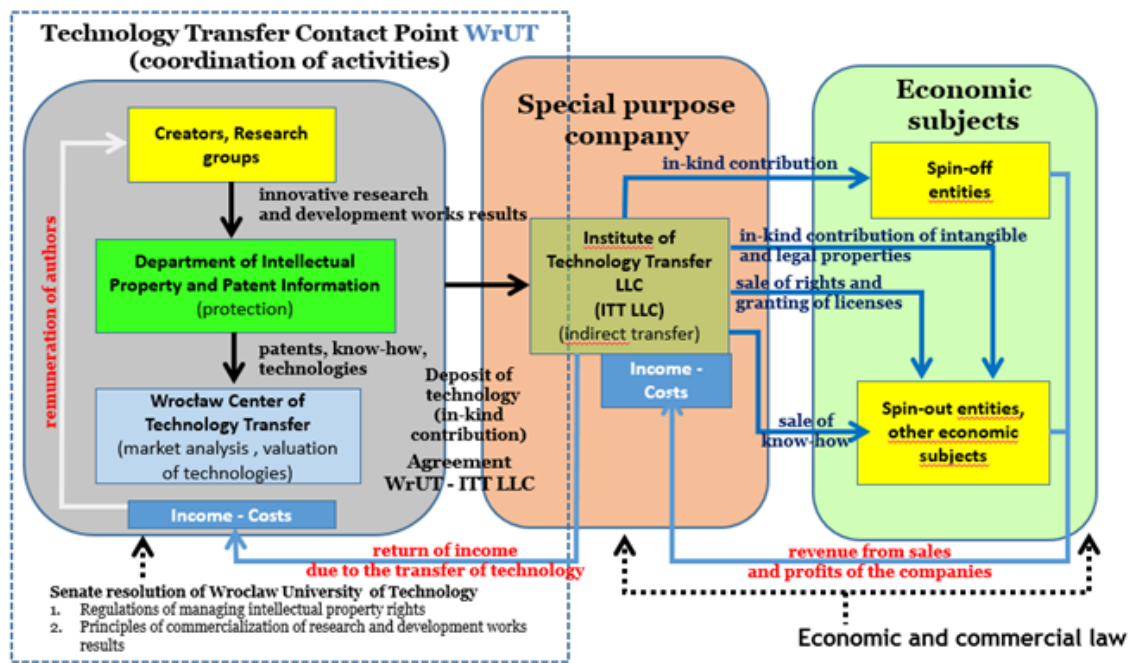


Figure 3. Model of indirect commercialization process in WrUT (Path 1 and 2)

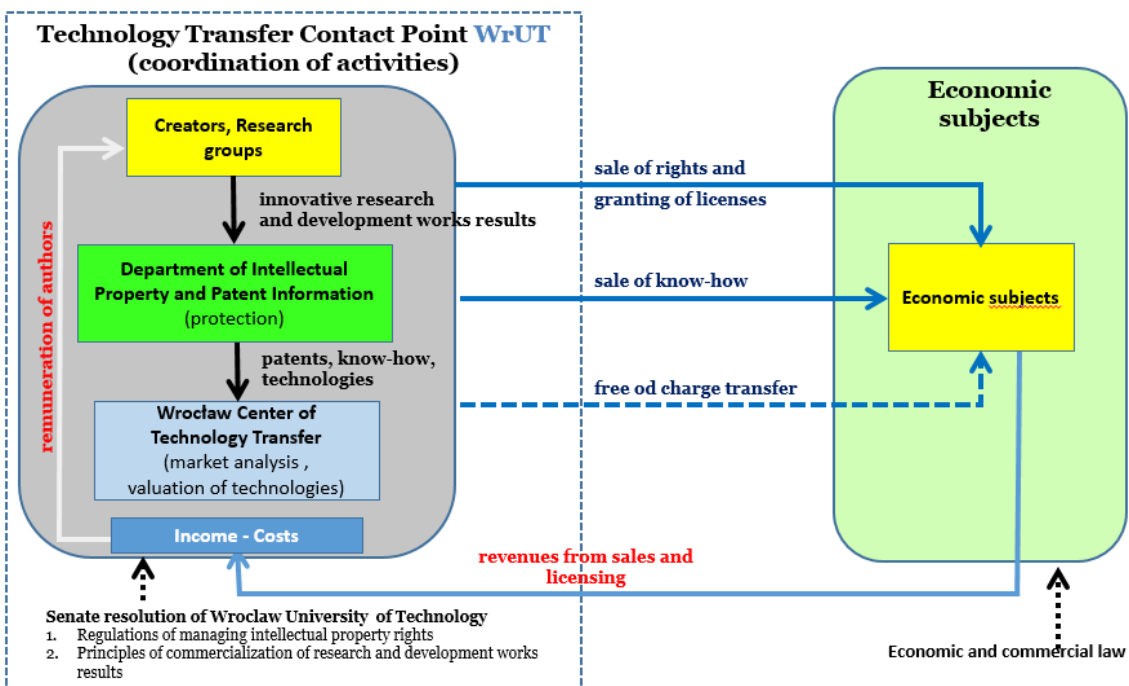


Figure 4. Model of indirect commercialization process in WrUT (Path 3 and 4)

ODINet - Online Data Integration Network

An innovative ontology-based data search engine

S. Pieroni, M. Franchini,
S. Molinaro
Institute of Clinical
Physiology, CNR
Pisa, Italy
{s.pieroni, m.franchini,
molinaro}@ifc.cnr.it

A. Greco, F. Pitto
Sistemi Territoriali S.r.l.
Cascina (Pisa), Italy
{a.greco, f.pitto}@sister.it

M. Toigo
Simurg Ricerche snc
Livorno, Italy
m.toigo@simurgricerche.it

L. Caterino
Rete Sviluppo S.C.
Firenze, Italy
caterino@retesviluppo.it

Abstract— Along with the expansion of Open Data and according to the latest EU directives for open access, the attention of public administration, research bodies and business is on web publishing of data in open format. However, a specialized search engine on the datasets, with similar role to that of Google for web pages, is not yet widespread. This article presents the Online Data Integration Network (ODINet) project, which aims to define a new technological framework for access to and online dissemination of structured and heterogeneous data through innovative methods of cataloging, searching and display of data on the web. In this article, we focus on the semantic component of our platform, emphasizing how we built and used ontologies. We further describe the Social Network Analysis (SNA) techniques we exploited to analyze it and to retrieve the required information. The testing phase of the project, that is still in progress, has already demonstrated the validity of the ODINet approach.

Keywords-Data search engine; domain ontology; semantic web; social network analysis.

I. INTRODUCTION

The Online Data Integration Network (ODINet) project is a Research and Development Project, approved within the Italian Regional Operational Programme having as the main objective the "Regional Competitiveness and Employment" through the 2007-2013 European Regional Development Fund. The project involves the prototypal implementation of an innovative semantic search engine (SSE) able to catalog numerical data in an ontological graph, to extract from data the information most relevant to the user requests using Social Network Analysis (SNA) algorithms and to return that information in a highly usable way.

The literature review indicates that there are multiple proposals for SSE, but none of them is specialized in finding information contained in alphanumeric datasets related to the user's needs. On the other hand, there are some search engines for numerical datasets, such as Quandl (<https://www.quandl.com>) and datahub.io (<http://datahub.io>), but it seems they are neither based on a semantic knowledge base, nor they employ SNA techniques.

Therefore, the ODINet's design is based on these prerequisites and its goal is to demonstrate the benefits of the combined use of these innovative principles.

The application domain is connected to the social, economic and health fields, in order to cover most of the

available data held by public bodies in the Italian context. Moreover, the three domains are closely linked to one another and offer the opportunity of cross-sectional cognitive investigations, through the identification of ontologies that describe interconnected concepts and links among various topics. Within each area, a kernel ontology has been designed for automatic and manual indexing purposes, bringing direct support to the SSE and enhancing retrieval accuracy. We further developed a data harvest component able to extract data from the web, interfacing to open data portals of Italian public administration.

We indexed those datasets together with the thematic ontologies, building a Search Graph that constitutes the main support for the Search Engine. This component, that is available as a web service, exploits well-known algorithms based on SNA properties, such as the centrality of nodes and the clusterization factor, in order to identify those datasets that are more related to the search query inserted by the user.

The search engine performs a semantic search on the graph: the semantic relations of our ontologies are enriched with two further procedures able to identify concepts in distinct ontologies that are semantically related one another. Finally, the identified datasets semantically connected to the user's query are returned by the web interface. A diagram describing the overall organization of ODINet platform is shown in Figure 1.

The focus of this paper is primarily the description of the semantic component of our platform, emphasizing the ontology-building process and how the ontologies have been used to build an index in form of a graph. We further present an in-depth description of our search engine component, explaining the whole process from a search query inserted by a user to the final results returned by the search engine.

The rest of this paper is organized as follows. Section II describes the knowledge resources and the ontology building process. Section III describes the technological platform and Section IV shows the main results. A final discussion closes the article.

II. KNOWLEDGE RESOURCES AND ONTOLOGY BUILDING

Building specialized ontologies from scratch through the support of domain experts' knowledge, requires a huge effort of conceptualization and a long editing time [1], especially in complex domains. Therefore, our choice was to rely on existing standard resources starting from EuroVoc [2], a

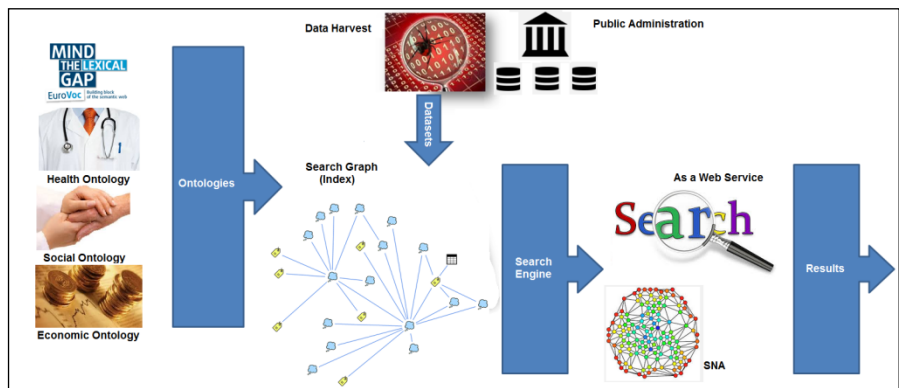


Figure 1. Overall organization of the technological platform

multilingual, multidisciplinary thesaurus, managed and maintained by the European Union’s Publications Office, which moved forward to ontology-based thesaurus management and semantic web technologies as Simple Knowledge Organization System (SKOS) conformant to W3C recommendations. EuroVoc has been widely used for classification software and indexers development [3], therefore starting from the SKOS [4] version of EuroVoc and relying on other standard resources, three specialized ontologies have been developed and linked in order to support the SSE. Since the ODINet's main objective is to access and classify a large amount of data and to present it to a wide range of users, EuroVoc has been chosen as it covers an exhaustive set of fields related to the activities of the European institutions, as shown in Figure 2. Therefore, in the initial phase of our work, the ontologies have been projected to have a wide horizontal spectrum, rather than a vertical one, using a top-down approach aimed to develop precise definitions of high-level concepts, and postponing to the validation phase a bottom-up approach necessary for analytic use-cases. After a deep analysis of the various sectors, we have chosen the ‘Social Questions’ domain as the core resource for both social and health ontologies. Actually, this domain copes with various relevant topics for the ODINet project as health, family, migration, demography, social framework and affairs, culture and religion and social protection.

Regarding the economic domain, we have chosen the sectors Economics, Trade, Finance, Business and Competitions, Employment and working conditions, Industry as core resources.

For the ontology editing, we adopted Protégé [5], which provides a conceptual development environment and an interactive graphical tool for the design and implementation of ontologies. Because of the project implementation and validation is conducted on data provided by Italian public institutions, the Italian version of EuroVoc sectors mentioned above, was transferred into Protégé creating one class, both for each sector and for each micro-thesaurus. In order to respect the original hierarchy, every micro-thesaurus has been linked to its broader/narrower terms through the native Protégé property SubClassOf. In coherence with EuroVoc specification, the relations hasBroader, hasNarrower and isRelated have been implemented inside Protégé. The annotation process has been mainly driven by the SKOS definitions of preferredLabel, hiddenLabel, seeAlso, isDefinedBy. Some overlapping concepts were present in the three domain ontologies. In the merging phase, this was addressed by means of SKOS's mapping properties used to define links between concepts in different ontology schemes, as exactMatch, closeMatch, narrowMatch, broadMatch and relatedMatch properties.

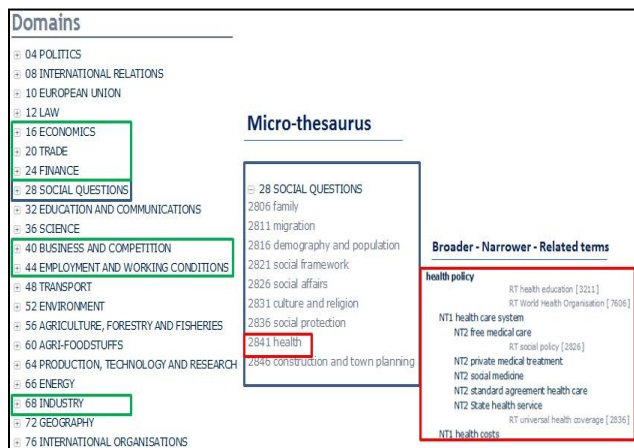


Figure 2. EuroVoc main structure

A. Health domain ontology

The main objective of Health domain ontology is to provide information about chronic diseases, with focus on cardiovascular disease and address 1) general questions coming from citizens, 2) specific questions coming from health system actors. In addition to EuroVoc, other specialized resources have been accessed in order to meet specific domain requirements, in particular the Unified Medical Language System [6] meta-thesaurus, a repository of biomedical concepts and the Medical Subject Headings [7] thesaurus, often used in document indexing. Efforts have been made to migrate the EuroVoc ‘2841 health’ micro-thesaurus (i.e., health policy, health care profession, illness, medical science, nutrition, pharmaceutical industry) and a part of MESH Heading related to the Cardiovascular Diseases ‘Tree C14’ to the standard formal web ontology language OWL [8], that became a World Wide Web

Consortium Recommendation in 2009. This migration step was crucial to performing ontology editing through Protégé and creating an initial interoperable ontology file. Then, we identified several extensions, the main ones regard drugs and disease: the extension related to drugs implied the migration of the first two levels of Anatomical Therapeutic Chemical (ATC) classification system. The extension related to diseases implied the migration of the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). The annotation process was implemented with the support of MESH thesaurus. A final extension was carried out to model measures and indicators of health and quality of care in Tuscany, developed by Health Regional Agency (ARS Toscana). Relations have been established between concepts of the different sub-hierarchies, e.g., given a diagnosis as Acute Myocardial Infarction, the ontology provide links with indicated drugs and with related diagnosis and symptoms.

B. Economic domain ontology

In recent years, the economic domain has shown several points of contact and overlap with the themes of health and social services. This partly because of the negative climate resulting from the world economy crisis, with strategies on spending review, compression of costs of health care, the links between income (individual and family) and access to health and social services [9]. These issues represent crucial points in the agenda of policy makers, due to the structural traits of economic crisis and it will become even more important to allocate resources in an optimal way, working on recovery aspects based on a cognitive framework in which the economic aspects will necessarily communicate with the aspects more directly related to the supply of services.

The economic domain ontology has been based on EuroVoc and some extensions have been identified. Primarily, metadata and the ontology of the Linked Open IPA public network and cooperation [10]. This resource contains the index of the national Public Administrations and the index of companies wholly owned by public authorities or with majority public capital included in the consolidated statement of public administration, as identified by the National Institute for Statistics (ISTAT). Secondly, the ISTAT Statistical Glossary [11] that provides classification and definitions of indicators and statistical terms used in the most open databases available in Italy. The final ontology has provided extension of several EuroVoc main sectors: sector 16 Economics, sector 20 Trade, sector 24 Finance, sector 40 Business and competition, sector 44 Employment and working conditions and sector 68 Industry.

C. Social domain ontology

Starting from the preliminary need to define the contours of the "social" in accordance with the objectives of ODINet, we have made an initial assessment of the Eurovoc resource. The nature of the instrument created to classify texts and documents could restrict the construction of an ontology that will mainly serve to classify and correctly interpret statistical data. So, we also considered the classification system defined

by the United Nations Economic Commission for Europe (UNECE), used as a base to build the structure of the Statistical Data and Metadata eXchange (SDMX) guidelines for the thematic classification of official statistics. UNECE has developed a classification of activities of statistics production (Classification of statistical activities - CSA 2009) based on three levels, the first of which consists of 5 main domains 1) Demographic and social statistics, 2) Economic statistics, 3) Environment and multi-domain statistics, 4) Methodology of data collection processing, dissemination and analysis, and 5) Strategic and managerial issues of official statistics. Within the domain Demographic and social statistics, we selected the three sectors Population and Migration, Social Protection and Social Policy and other community activities; within the domain Environment and multi-domain statistics, we chose the sub-sector living conditions, poverty and cross-cutting themes. We then entered all the specific concepts used in the production of official statistics, identifying them from ISTAT and comparing with Eurostat articulation [12] [13] [14]. The definitions have been built with the aid of additional glossaries and thesauri, such as United Nation Common Database [15] and the International Statistical Institute Multilingual Glossary [16]. Subsequently, we have enriched the concepts and definitions drawn from the statistical glossaries and we have extended the theme of social protection by referring to the European system of integrated social protection statistics (ESSPROS), with its annotations and relations. Preferred and alternative labels and definitions have been incorporated by reference to EuroVoc and to terminology and correlations themes taken from the "Thesaurus for the Social Sciences" developed by the Leibniz Institute for the Social Sciences.

III. TECHNOLOGICAL PLATFORM

In this section, we describe ODINet technological platform main components.

A. Data Harvesting

We designed and developed a complex module able to interface with existing portals and to find accessible datasets in the web. The module, can be periodically scheduled and manages to automatically import meta and content information from a wide variety of formats, such as CSV, XLS, MDB, DBF, SHP and RDF. Data were drawn mainly from *I.Stat* (database of statistics produced by ISTAT), *dati.toscana.it* (an open data platform developed by the Tuscany Region) and *dati.gov.it* (an open data portal developed by the Italian Ministry for Public Administration and Innovation) portals. Stakeholders who joined the project and provided data for the validation scenario are *ARS Toscana* (the Regional Health Agency), *IRPET* (the Regional Institute Planning Economic of Tuscany) and *Rete Osservatori Sociali Regione Toscana* (a Social Observers Network in Tuscany).

B. Data Indexing

A search graph was built as main support for the search engine by combining the concepts identified during the

ontology-building process and adding relationships between them according to ontologies' predicates. Since the graph was sparse, two procedures were designed and implemented to improve its semantic information. Finally, the data component was added to the graph and linked to the concepts contained in it. In such a way, the complete domain knowledge is summarized in a single graph, which can be used as the basis of the reasoning mechanism of the Search Engine to answer users' queries.

1) From the ontologies to the Search Graph

In order to produce an effective and useful search engine we built a search graph combining the previously described ontologies and the data harvested. Having a weighted graph representing our model of knowledge is a key element in our project, since it allows to answer users' search queries by browsing the graph through well-known algorithms based on distance metrics, centrality of nodes and clustering coefficient, that were partially developed and tested in a previous research project [17]. We decided to store into different graph entities the concepts corresponding to the various concepts identified inside ontologies, as well as datasets (corresponding to collections of data organized in a single table), literals (corresponding to labels associated to concepts or keywords associated to datasets) and categories (corresponding to generic themes grouping datasets from the same semantic area). Each concept has at least one associated literal, corresponding to its name, but can also be associated with multiple literals, corresponding to alternatives synonyms for that concept. For example, the concept identified by the url "http://eurovoc.europa.eu/5565" has as its preferred label the literal "cyclone" but has three further associated literals, namely "hurricane", "tornado" and "typhoon". In the current graph we have ~10K concepts, ~24K literals, 22 categories and ~9K datasets. All these entities are the nodes of our graph. Afterwards, we introduced into the graph several relationships between nodes. The relationships are established on the basis of the ones identified in the domain ontologies and SKOS predicates. Each type of relationship has a specific weight in the search engine process. In the current graph we have ~100K relationships.

2) Enriching the Search Graph with semantic information: concept-concept matching.

Analysing the search graph we have described so far, we noticed that it was extremely sparse, as we had few edges compared to the number of nodes. Since the activity of finding connections between concepts that belong to different ontologies takes a long editing time even to domain experts, we explored automated methods for discovering those links. Our main goal was to identify concepts and literals semantically connected to one another: in this way, given a search string it would be possible to identify a group of concepts semantically connected to the string searched by the user, identifying not only datasets containing the string, but also the ones connected to a concept semantically related to it. To identify entities semantically connected to one another we implemented two separate procedures named *WikiSimilarityDistance* and *WikiConceptConnection*.

a) WikiSimilarityDistance

This procedure produces an estimate of the semantic distance between two concepts through an approximation of the Google Similarity Distance: this measure calculates the semantic similarity between two words (or two sentences) on the basis of the number of pages indexed by Google in which the two words (sentences) appear together, in relation to the number of pages in which they appear singularly. The formula to calculate the (Normalized) Google similarity Distance (NGD) is the following:

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - \min\{\log f(x), \log f(y)\}}$$

where $f(x)$ and $f(y)$ are the number of hits for words x and y respectively, $f(x,y)$ is the number of pages in which both x and y occur and M is a parameter to take into account the order of magnitude of the number of results. Since Google APIs to access a search result are not free, we approximated this measure by indexing in Solr [18] a dump of the whole Italian Wikipedia and performing search queries against the indexed Wikipedia pages for concepts and literals that are neither too general nor too specific. Couples of concepts with an NGD value below a given threshold are said to be semantically correlated. We managed to identify, through such a procedure, ~10K semantic correlations between entities, enriching our search graph with semantic connections of different weights, according to their NGD values.

b) WikiConceptConnection

Through this procedure we managed to find further semantic associations between couple of entities exploiting the items in the "See Also" section of Wikipedia pages. The procedure associates concepts and literals in our graph to Wikipedia pages with the same name, then it parses those pages and retrieves the links in the "See Also" section, which contains a list of pages that are correlated to the current one. If a concept (or a literal) with the same name of one of the pages in the "See Also" list exists, a new edge between the two correlated entities is inserted into the graph. Through such a procedure we managed to add about 3.5K further relationships to the Search Graph.

3) Inserting datasets into the Search Graph and dataset-concept matching

For each dataset, we decided to import its title (so as to identify it), its keywords (literals to which the dataset is connected) and its description (so as to get more information about its content). Then, we connected each dataset to the concepts expressed in it. Since semantic correlations between concepts were already stored into the graph, we decided to perform a purely-syntactic matching between datasets and concepts and to index all the information about a dataset using Solr [18]. We performed a Solr query for every concept in our graph, searching its associated preferred label in the title, in the keywords and in the description of the indexed datasets. A new relationship was created between a concept and each dataset in which the concept was found to be expressed. We further assigned different weights to each of those relationships according to where and how many times the concept was found in the dataset. One of our

assumptions was, for example, that if a concept was found in the dataset’s title the connection with the dataset is stronger than if it was found in the description. Finally, we ran a similar procedure to match datasets with literals whose names were not the same as the one of the concepts. In total we managed to add ~20K further relationships to the Search Graph.

4) Visual representation

The visual representation of the Search Graph is shown in Figure 3. A zoom-in snapshot of the graph is depicted on the left side: the cloud icon identifies a concept, the table one represents a dataset, the one similar to a label represents a literal and the green one identifies a category. A zoom-out snapshot of part of the graph is shown on the right.

C. Data Search Engine

This paragraph explains how the search process works, starting from a search query entered by a user up to the final output returned by the Search Engine.

1) Full-text Search

In this step, we aimed to find concepts, literals and datasets that are related to the search string entered by the user. A SQLServer [19] query, performed a full-text search to find entities containing all the meaningful words (articles, prepositions and common words are deleted) entered by the user. A list of datasets containing all the search words is memorized and returned in the final results subsequently. If neither concepts nor literals are found, the datasets obtained with the query is directly returned as the final result and the procedure ends.

2) Page Ranking and Semantic Propagation with SNA

a) 1st propagation

In this step, we aimed to find concepts and literals semantically connected with the ones returned by the previous phase. For this purpose, we implemented a variant of the PageRank algorithm [20] to identify concepts and literals most strongly connected with the preceding ones. This algorithm executes a propagation on the search graph by taking into account the strength of a link to an entity (represented by its associated weight) in order to determine a rough estimate of how important the entity is. This step produced a list of concepts and literals semantically connected with the search query.

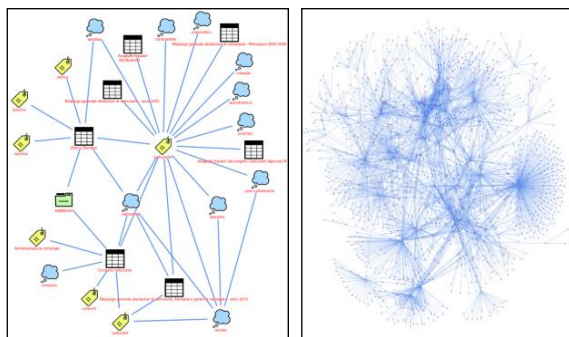


Figure 3. Visual representation of the Search Graph

b) 2nd propagation

We used the PageRank algorithm to find datasets directly connected to literals and concepts returned by the previous step. At the end we got a list of datasets connected to concepts related to the search query.

3) Final Results

The final result returned to the user is the union between the datasets identified with the full-text search and those returned at the end of the second propagation phase. The datasets are ordered by their rank value, a measure calculated during the search process that estimates the dataset’s relevance for the search query. The flowchart of the search engine process is shown in Figure 4.

IV. RESULTS

As a first result, the ontologies may function independently and by themselves as separate knowledge basis. They provide about 4000 distinct concepts, corresponding relations and labels for synonyms.

In order to validate the overall research, we defined some contexts and use cases based on real issues identified with stakeholders. The *Caring for the elderly* use case is shown in Figure 5. The list of concepts found by the semantic search is Zonal user rate for elderly home care, Health and social care for the elderly, Taking care of elderly people for the health service, Caring for the elderly, Elderly person, Gerontology, Elderly people, Family solidarity, Pension scheme, Care allowance, Retired person, Home help, Facilities for the disabled, Older worker.

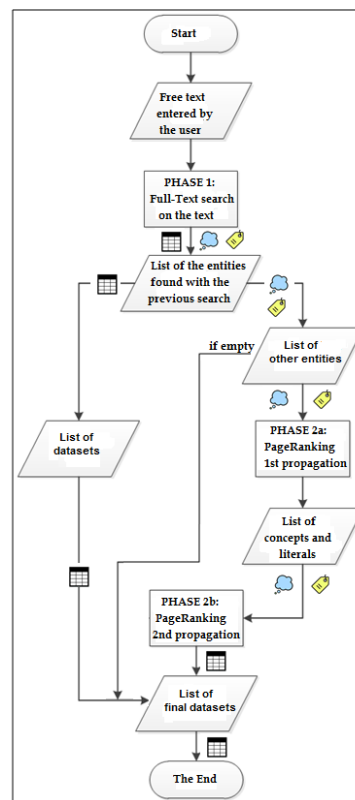


Figure 4. Flowchart of the search engine process

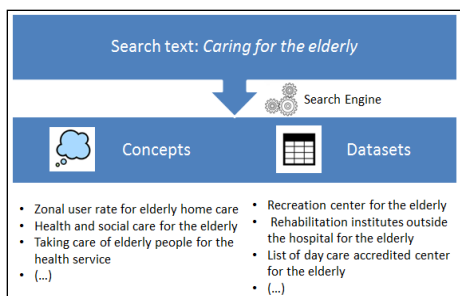


Figure 5. Use case *Caring for the elderly*

The top five datasets returned are Recreation center for the elderly, Rehabilitation institutes outside the hospital for the elderly, List of day care accredited center for the elderly, Social services for the elderly and List of residential structures for the elderly. A dynamic web interface which displays the search results organized in semantic clusters is under testing: the user will be able to see graphical representations of the results, to select only those datasets that are more strongly connected with his search (making a dynamic disambiguation) and to navigate the associated graph.

Our platform has a number of strengths. Firstly, our search engine is both reliable, being based on well-known SNA algorithms working on graphs, and innovative, being the ontologies and the numerical datasets included in a single tool that constitutes the knowledge base of the whole system. Further, the knowledge representation's component and the search engine's module are decoupled, guaranteeing a high level of reuse and adaptivity. In fact, with proper changes to the underlying ontology, the tool can be used in completely different contexts with respect to the current ones. Moreover, the harvest procedure can be periodically scheduled and manages to automatically import data and to index the newly found datasets.

Another innovative aspect of our tool is to extend the semantic component of the system to the whole knowledge base. Our SSE is based on an ontological graph of entities, among which there are generic concepts and datasets, semantically connected to one another. This fact lets the SSE to identify datasets relating not only to the search query but also to concepts semantically related to those contained in the search text. Lastly, differently from other systems, the user can directly interact with the tool, making a dynamic disambiguation on the results and identifying those datasets that are more strongly connected to his search.

V. DISCUSSION AND FUTURE DEVELOPMENTS

In this paper, we have mainly presented the semantic component of the ODINet project (<http://www.odinet.sister.it>), which provides an innovative technological framework for data search engine.

Specifically, we have first created a unified ontology which models the conceptual hierarchies that describe the major aspects of social, economic and health fields and the main connections between them. We have designed and developed a complex module able to interface with existing

portals and to find accessible datasets in the web. A search graph was built as main support for the SSE by combining the concepts identified during the ontology-building process and adding relationships between them according to ontologies' predicates. A dynamic web interface was developed to display the search results in an intuitive way, allowing the users to play a part of the answers and perform disambiguation.

The validation phase of the project is still in progress. Different stakeholders offered validation scenarios on thematic portals, correspondent to ODINet domains. The tests carried out so far have shown that our platform achieves higher performances on those portals than on generic datasets: on a number of test cases performed so far, we found that more than 80% of the datasets returned by the SSE are actually related to the search query. This percentage rises to about 90% for queries related to the ODINet's domain ontologies. In fact, our search graph was built combining thematic ontologies containing specific concepts that belong to those thematic areas.

Our system also has some weak points, among which we found that connections between concepts and datasets are not always strong enough. This mainly depends on the level of detail of the meta information associated to the datasets. Further, ODINet ontology is in Italian. However, as EuroVoc and every used resource are also available in other languages, an English version can be provided with moderate effort in the future.

Moreover, in the next few months, the technologies implemented by the ODINet data search engine will be integrated in "StatPortal Open Data" (<http://www.opendata.statportal.it>), an open source platform for developing open data portals.

ACKNOWLEDGMENT

This work is supported by the Italian Tuscany Regional Operational Programme - 2007-2013 European Regional Development Fund (ODINet project D57112000710007).

REFERENCES

- [1] S. Gedzelman, M. Simonet, D. Bernhard, G. Diallo, and P. Palmer, "Building an Ontology of Cardio-Vascular Diseases for Concept-Based Information Retrieval", *Computers in Cardiology*, vol. 32, 2005, pp. 255-258.
- [2] Office for Official Publications of the European Communities, "Thesaurus EUROVOC" Annex to the index of the Official Journal of the EC, Luxembourg Publications of the European Communities, 1995.
- [3] R. Steinberger, M. Ebrahim, and M. Turchi, "JRC EuroVoc Indexer JEX - A freely available multi-label categorisation tool" *Proc. LREC'2012*, May 2012, pp. 798-805.
- [4] N. Ivezic, A. Farhad, K. Khosrow, and B. Kulvatunyou, "Ontological Conceptualization Based on the Simple Knowledge Organization System (SKOS)", *Journal of Computing and Information Science in Engineering*, doi:10.1115/1.4027582, vol. 14, issue 3, May 2014, pp. 11.
- [5] M. Horridge, "A Practical Guide To Building OWL Ontologies Using Protege 4 and CO-ODE Tools", *The University Of Manchester Edition 1.3*, March 2011

- [6] O. Bodenreider, "The Unified Medical Language System (UMLS):integrating biomedical terminology", *Nucleic Acids Res.* doi:10.1093/nar/gkh061, vol. 32, 2004, pp 267–270
- [7] W. D. J. Stuart, J. Nelson, and B.L. Humphreys, "Relationships in Medical Subject Headings (MeSH)." National Library of Medicine, Bethesda, MD, USA, 2002. Available from: <http://www.nlm.nih.gov/mesh/meshrels.html> [retrieved: 12, 2014]
- [8] D. L. McGuinness, and F. van Harmelen, "OWL Web Ontology Language Overview". W3C Recommendation, February 2004. Available from <http://www.w3.org/TR/2004/REC-owl-features-20040210/> [retrieved: 12, 2014]
- [9] World Health Organization on behalf of the European Observatory on Health Systems and Policies, "Health, health systems and economic crisis in Europe: impact and policy implications", 2013. Available from <http://www.euro.who.int> [retrieved: 12, 2014]
- [10] Italian Public system of connectivity and cooperation, "Guidelines for semantic interoperability through linked open data", 2012 Agency for Digital Italy
- [11] M. Frustaci, "Glossary Economic-Statistical Multilingual". Italian National Institute for Statistics ISTAT, Doc 8/2004, Available from <http://www.istat.it> [retrieved: 12, 2014]
- [12] European Commission - Eurostat, "Eurostat: The Statistic Explained Glossary". Available from http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Thematic_glossaries [retrieved: 12, 2014]
- [13] European Commission - Eurostat, "RAMON Eurostat's Metadata Server". Available from <http://ec.europa.eu/eurostat/ramon/index.cfm> [retrieved: 12, 2014]
- [14] European Commission - Eurostat, "Coded - The Eurostat concepts and definitions database". Available from http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL_GLOSSARY&StrNom=CODE D2&StrLanguageCode=EN [retrieved: 12, 2014]
- [15] United Nations Statistics Division, "United Nations Common Database – Methods and classifications". Available from <http://unstats.un.org/unsd/methods.htm> [retrieved: 12, 2014]
- [16] The International Statistical Institute ISI, "The multilingual ISI glossary of statistical terms". Available from <http://isi.cbs.nl/glossary/> [retrieved: 12, 2014]
- [17] M. Baglioni, S. Pieroni, F. Geraci, S. Molinaro, M. Pellegrini, and E. Lastres, "A New Framework for Distilling Higher Quality information from Health Data via Social Network Analysis". 13th International Conference on Data Mining (ICDMW.2013) IEEE, December 2013, pp. 48-55, DOI 10.1109/.142.
- [18] T. Grainger and T. Potter, "Solr in Action". Manning Publications, 2014.
- [19] R. Mistry and S. Misner, "Introducing Microsoft SQL Server" 2014. Microsoft Press, 2014.
- [20] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", Technical report, Stanford Digital Library Technologies Project, 1998

If Experience is Worth, How Experts Behave in a Manga Case

Satoshi Takahashi

Department of Computational Intelligence and Systems
Science
Tokyo Institute of Technology
Kanagawa, Japan
e-mail: takahashi.s.bh@m.titech.ac.jp

Toru B. Takahashi

Department of Management Science
Tokyo University of Science
Tokyo, Japan
e-mail: takahashi.toru@rs.tus.ac.jp

Atsushi Yoshikawa

Department of Computational Intelligence and Systems
Science
Tokyo Institute of Technology
Kanagawa, Japan
e-mail: at_sushi_bar@dis.titech.ac.jp

Takao Terano

Department of Computational Intelligence and Systems
Science
Tokyo Institute of Technology
Kanagawa, Japan
e-mail: terano@dis.titech.ac.jp

Abstract— To develop and put in place effective training methods in business, discovering how experts and novices differ is important. However, it is difficult to measure these differences in an actual work environment. In this study, that problem is resolved by using a manga case, in which business scenarios are illustrated in a distinctive Japanese comic book form. The characteristics of the manga case provide the reader with a hypothetical business situation experience. In the study, experts and novices were placed in a simulated business situation using a manga case and were asked to evaluate the leadership skills of one of the characters. From those evaluations and participant interviews, two differences were observed: 1) The experts paid attention not only to what the leader did and said, but also to background information, such as what staff members did and said and the description of the office; 2) The breadth of information considered was larger for the experts than the novices in terms of “scale” and “time.” “Scale” refers to the number of individuals and organizational factors, etc., considered as they weigh the effects of events and the cause-and-effect relationships. “Time” refers to the chronological spans considered when weighing the effects of events and the cause-and-effect relationships.

Keywords—business; expert assessment; manga case; case method; leadership.

I. INTRODUCTION

For companies, the speed of technological change continues to accelerate. One effect of that acceleration is that businesses are searching for more effective training methods [1]. To develop and put in place effective training in business, discovering how experts and novices differ is important. Once such differences are understood, it will be possible to conduct research regarding an educational approach that can effectively address these differences. In this study, these differences mean the different ways experts and novices behave and perceive work-related situations.

One obstacle, however, is that in reality it is difficult to measure the behavior and perception of experts and novices as they engage in work. As a result, in past research on business experts, accomplishments, years of service, and psychological measures have been studied [2]. This makes it difficult to use such results to develop more effective business training methods.

However, studies on experts in sports and in education have been conducted by Kato et al. and Chi et al., respectively [3] [4]. Kato et al. had participants simulate sports play and, by measuring changes in line of sight, discovered that there were differences between experts and novices. Chi et al. had participants solve a physics problem while verbalizing their thought process (Think-aloud protocol). Their study found that there were differences between the experts and the novices in how they perceived the problem. In both studies, a simulated context was devised for the experiment to discover the differences between experts and novices and measure behavior and perception.

In this study, a manga case was used to simulate business situations in order to discover expert-novice differences in behavior and perception.

Manga cases are designed for educational purposes and present typical business situations in a comic book format [5] [6] [7] [8]. Manga cases can be defined as case materials in a comic book form. They can provide simulated business experiences for experiment participants via the illustrated format. Further explanation about manga cases is provided in Section II.

In this study, Section I is the introduction; Section II provides an explanation of manga case characteristics; Section III explains the experiment’s methodology; Section IV presents the results ; Section V presents the discussion; and, Section VI presents the conclusion.

II. THE MANGA CASE

This Section describes the characteristics of manga cases and how they are used. A manga case is a genre of educational material that presents scenes and story lines important to business. Manga case topics and usage are similar to those found in case method. To date, we have created manga cases to illustrate topics, such as worker-management conflicts and negotiations with client companies.

Compared to case materials in which the business scenario is described in words, manga cases describe the scenario through a sequence of panels that make up a comic strip. In a manga case, information is embedded in the material via use of characteristic comic techniques of expression; that is, facial expressions and words of characters, and background drawings of the office furniture and cityscapes. The reader discovers pieces of information embedded in the educational content and interprets them to build a hypothesis. In this way, the reader can come close to experiencing a real business situation.

A few examples of embedded information and hypothesis building are given below from “Website Flaming!” used in this study.



Figure 1. Information embedded in expressions, words, and actions.

A. Communication through Facial Expression, Words, and Action

An example of communication through facial expressions, words, and action is shown in Figure 1, depicting a meeting at a startup company. In scenes from that meeting, elements, like the characters’ words and facial expressions, communicate the nature of the relationships among the characters and how they share information:

Woman (Systems Engineer): “So, in order to get stable profits for the company...”

Man 1 (Second in command): “Hold on, if you think about VC...”

Man 2 (President): “OK, OK”

The reader can see from the woman’s expression and from what she says that she is unhappy with a recent organizational decision; and, gather from the president’s

expression and words, that he is the type of character who wants the meeting to proceed amicably.

B. Communication via Background Depictions

Figures 2 and 3 illustrate communication via background depictions. Figure 2 depicts a whiteboard in the startup company’s office. From the schedule written on it, the reader can interpret the method and status of information sharing within the company. Specifically, at this company, transaction negotiations with a major client and a briefing session for a venture capital Board are scheduled to take place in the near future, but those events are not written on the schedule. From that information, the reader can infer that management information in this company is not shared with the staff.

Figure 3 depicts the startup company’s office space. From this scene, the reader can extrapolate the office size, equipment, the number of employees, the amount of capital invested in equipping the office, the working hours, etc.

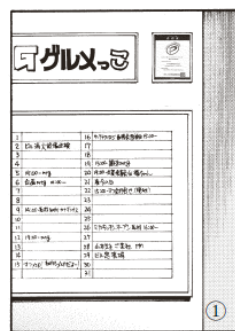


Figure 2. Background depicting a whiteboard.



Figure 3. Background depicting the office.

C. Information Discovery and Hypothesis Building

The reader gathers and synthesizes information from panels, such as those shown in Figures 1, 2, and 3, and builds a hypothesis about what is happening. An example of this hypothesis development follows.

Hypothesis: From the investment in equipment, the number of employees, and earnings issues, it appears that the company’s finances are not healthy. On the other hand, in the near future, there is going to be a meeting with the venture capital Board. That means that tonight’s negotiations with the major client company need to be successful. However,

Man 2, the president, has not shared that information with the staff, and is trying to resolve the problem alone.

III. METHODOLOGY

This Section explains the details of the methodology used in the study. Business experts and novices were recruited as participants, provided with the manga case, and asked to respond to a set of problems. After responding to all the problems, they were then interviewed. The interviews consisted of questions about the participants' thinking and what they paid attention to as they responded to the problems. More details are provided below regarding the manga case, the participants, the problems, and the interview questions.

A. The Manga Case

For the experiment, the manga case used was "Website Flaming!" which explores the case of an IT startup company. The main characters are Mr. Tanaka, the president of the startup company, Ms. Isaka, the company's systems engineer, and Mr. Chiyokura, the company's second in command. Other characters include the company's other staff and the staff of the major client company. The general story line is that the website of the startup company is about to go up in flames and though they have made various attempts to prevent this from happening, it eventually catches fire and takes down the company's servers.

B. The Participants

The participants in this study consisted of four university students as business novices (N1, N2, N3, N4) and four company employees as experts (E1, E2, E3, E4). The experts were selected referencing the 10-year rule, and, therefore, had 10 years or more of working experience [9].

C. The Problems

The following three problems were set up related to leadership skill evaluation:

Problem 1: Choose 20 or more panels that are relevant to evaluating Mr. Tanaka's leadership.

Problem 2: Use the provided leadership evaluation scale to assess Mr. Tanaka's leadership.

Problem 3: Based on your evaluation in problem 2, select additional panels to those you chose in problem 1 that are relevant to your assessment.

There are two reasons why the evaluation of leadership was chosen as the theme for the problems. First, in business, leadership evaluation is considered a necessary skill regardless of business type. Second, knowledge specific to a particular business is not a necessary factor in leadership evaluation. "Knowledge specific to a particular business" refers to the knowledge learned after joining and working at

a company. The knowledge about financial operations in a startup company would be an example of this. Basing a problem on this type of knowledge would likely introduce bias into the results regarding the differences between working participants and student participants, by including an effect from such knowledge.

The reasoning behind problem 2 was to prevent bias from the differences in the participants' perceptions regarding leadership. By using a common measure of leadership to evaluate Mr. Tanaka's skills, participants were forced to somewhat share the same leadership standards. More specifically, the initial selection of panels in problem 1 was created as a separate step to enable an examination of how exposure to problem 2 might affect expert and novice participants differently. The results of that analysis are omitted here since they fall outside of the purpose of this study.

For problem 2, the Performance-Maintenance (PM) Leadership Scale was employed because it is standard leadership evaluation scale. The scale evaluates leadership based on performance and maintenance items [10]. Although there are a few versions of the PM Leadership Scale, in this study the scale for top management leadership was used [11]. The top management scale has 77 items. Part of the scale is shown in Table I. For each PM evaluation item, the participants were to choose one out of four ratings: "good," "bad," "seen in some panels but no judgment is possible for the whole story," and "not seen in any panel."

TABLE I. PM LEADERSHIP SCALE(TOP MANAGEMENT LEADERSHIP)

Item #	Evaluation Item
1	Demands the accomplishment of priority objectives.
2	Requires that decision-making has an objective and quantitative basis.
3	Once an objective has been set, it is pursued, even when difficulties are encountered.

D. The Interviews

Interviews were performed after all the problems had been answered. In the interviews, participants were asked about Mr. Tanaka's leadership evaluation and the reasons for their choices when answering problems 1 and 3.

IV. RESULTS

This Section presents the results obtained from the methodology explained in Section III, followed by a discussion. First, the answers to problems 1 and 3 were quantitatively analyzed. Next, the interview answers were qualitatively analyzed.

A. Quantitative analysis

This Section presents a quantitative analysis of the answers to problems 1 and 3. Figures 4 and 5 show the results of problems 1 and 3. The horizontal axis of Figure 4 shows each participant and the vertical axis shows the

number of panels selected. The number of panels related to Mr. Tanaka’s words and actions was tabulated separately from those panels without his words and actions (referred to below as “background information panels”). In Figure 5, the ratio of background information panels to the total number of panels selected is shown on the vertical axis against the horizontal axis showing the novices and experts.

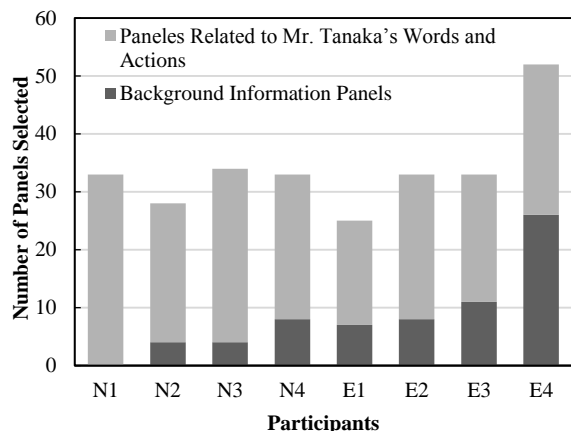


Figure 4. Number of panels selected by participants

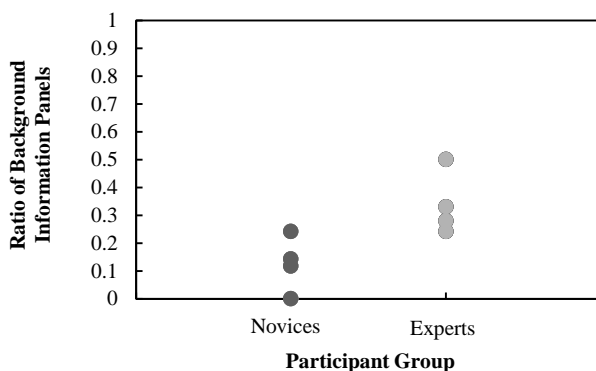


Figure 5. Ratio of background information panels.

In Figure 4, a difference can be seen in the number of background information panels that the novices selected, 0–8 (with an average of 4), compared to the experts’ selecting 7–16 (with an average of 13). In addition, in Figure 5, a difference can be seen in the ratio of background information panels to the total selected, 0.000–0.242 (avg. 0.126), by the novices, compared to 0.242–0.500 (avg. 0.338) for the experts.

The ratios of background information panels to the total number of panels selected were tested for variance in the means of the novice and expert groups. First, in order to test for equal variance, an F-test was performed. The result of the F-test showed that the variance between novices and experts was not significant with a one-tailed $P=0.41 > 0.05$ result. Next, a t-test was performed assuming an equal variance for

the two samples. The t-test showed a significant difference in the means between the novice and expert groups with two-tailed $P=0.03 < 0.05$ result.

B. Qualitative Analysis

In this Section, the qualitative analysis is presented. The analysis consisted of a repetitive process of: 1) segmenting of the interview results; 2) coding; and 3) thematic grouping. The results showed two key themes in the interviews: “scale” and “time.”

“Scale” is the scale of items being considered at the time of the interview. The four items discovered, from a small to a large scale, were: Mr. Tanaka’s personal attitude and nature, Mr. Tanaka’s relationships with staff, Mr. Tanaka’s relationship with the outside world, and the overall organization.

“Time” is the chronological relationship between the events in the story line and the events the participants were considering at the time of their comments. The three periods identified were: the past, during the story line, and the future. To clarify, “the past” refers to an event the participant considered had already happened before the story began. “The future” refers to an event the participant considered will happen after the story ends.

Against the themes of “scale” and “time,” the interview results are shown in Tables II and III. As can be seen in these tables, novice comments rarely related to the “overall organization,” to the “past,” or to the “future.”

The next few paragraphs give examples of expert comments, accompanied by the panels they refer to. These illustrate three themes with differences between the experts and novices: the “overall organization,” the “past,” and the “future.”

1) The “overall organization”

The first example is shown in Figure 6, a scene where employees are having a conversation. For this scene, one expert commented, “They are not on the same page,” implying that, the leader is neglecting to pay attention to enabling smooth communication among staff.

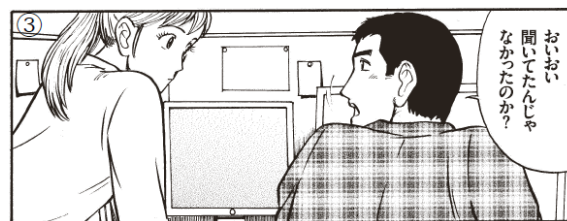


Figure 6. A conversation between staff members.

Second in Command: “Weren’t you listening?”

A second example is shown in Figure 7, a panel depicting a whiteboard in the office. Regarding this panel, one expert commented, “Even though it is year-end and there must be meetings with the venture capital Board and banks, that hasn’t been shared.” By this, he meant that, when

information important to the operation of the organization is not internally shared, a gap in awareness arises between the leaders and staff.



Figure 7. The whiteboard.

As in these examples, experts evaluated Mr. Tanaka's leadership skills based on information related to the entire organization. Elsewhere, some experts, noticing elements like the office equipment, threw doubt on the firm's balance in allocating funds. Additionally, there were comments pointing to the immaturity of the systemic response to the server problems.

On the other hand, novice comments were largely limited to Mr. Tanaka's individual behavior. The only exception was one comment stating that, "Good benefits" in response to a scene in which Mr. Tanaka is explaining about bonuses.

2) The "Past"

Figure 8 depicts a business dinner scene with a client in which the president's response to the situation is questionable. Regarding this scene, an expert commented that, "They didn't prepare for the meeting by working out an integrated approach internally beforehand." By this he meant, usually, before an external meeting, an integrated approach is worked out internally, including roles and conversation content. However, because Mr. Tanaka does not trust his staff, that probably never happened.



Staff member: "He should push the word-of-mouth side as we do that. Do they want us to simply do subcontract work?"

Figure 8. A lack of trust from staff.

In these comments, the expert is looking for cause-and-effect relationships from the past that preceded the story's timeline. Another expert made comments implying that, a strategy to deal with the server problems (in the story) should have been put together before the story, not during it.

On the other hand, novices tended to identify causes or measures only in actions or events within the story line and they had no comments on the past before the story's time line.

3) The "Future"

The first example of the future theme is depicted in Figure 9, in a scene in which Mr. Tanaka responds to a question from the client. In this scene, he replies that there are no problems, in response to the client's question regarding whether there have been problems with the website. Regarding this scene, an expert commented that, "If there really were problems, to respond that way is a bad idea," meaning that, if there were truly problems, there could be situations later on when it would be difficult to explain these to the client and the client's trust would be completely lost.

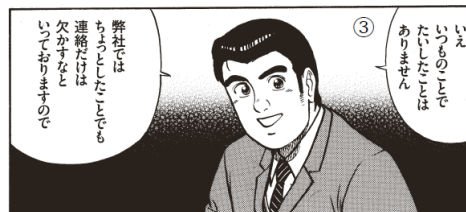


Figure 9. Response to a client.

Tanaka: "There's no problem. It's something that happens all the time. I always tell the staff to keep in touch even about the slightest things."

A second example, in Figure 10, is a scene of an internal meeting where Mr. Tanaka is giving an explanation about the business going forward. Regarding this scene, an expert commented that, "He's thinking about new business," by which he meant, Tanaka is thinking not only about the current business but about future business.



Figure 10. Tanaka's future business vision.

Tanaka: "So, we've started website management of 'word-of-mouth websites' in order to expand the company, and its finally started to achieve a little turnover this year. We're a long way from operating in the black on a year-to-year basis with just this project, and the management costs aren't cheap, but if you think about the spinoff business which can be derived from word-of-mouth, it's clear that it will grow. No, we have to make it grow."

In these examples, without limiting their evaluation to the results of actions within the story, the experts appreciated the possible impact on the company's future after the story ended.

In contrast, the novices tended to look for results and evaluations of Tanaka's actions only within the story line, and made no comments about the future beyond the story's time frame.

V. DISCUSSION

The qualitative analysis showed that when evaluating leadership, experts referred to a larger amount of background information, defined as information other than the words and

actions of the leader. Examples of this were conversations between staff and information written on the whiteboard.

The results of the qualitative analysis showed that the experts referred to information broad in scale and time in order to evaluate leadership skills. Being “broad in scale” meant that the experts focused not only on the leader’s actions and words but also on the situation of the overall organization. Being “broad in time” meant that the experts did not limit themselves to considering actions and results within the story, but rather found causes in past behavior (before the story) and also evaluated the impact of actions in the future after the story.

Synthesizing the results of the quantitative and qualitative analyses, thinking that was “broad in scale” seemed to result in the experts paying attention to background information, such as conversations between staff members in Figure 6 and the whiteboard in Figure 7. Additionally, thinking that was “broad in time” seemed to result in the experts paying attention to background information, such as a lack of trust from staff in Figure 8.

The results can be summarized as follows: It may be seen that the experts tended to consider information more broadly in terms of scale and time, and as a result, the ratio of background information they used was larger.

VI. CONCLUSION

This study used a manga case to identify differences between expert and novice assessments in the evaluation of leadership skills. The results of this study suggest two points regarding how experts and novices differ as they evaluate leaders. First, experts may tend to observe not only the actions and words of leaders, but also background information, namely, information, such as what staff members do and say, and how the office looks. On the other hand, novices may tend to depend more on observations of the leader’s actions and words. Second, the breadth of information taken into consideration by experts may be larger than that of novices in two ways, “scale” and “time.” “Scale” refers to the number of individuals and organizational factors, etc., considered as they weigh the effects of events and cause-and-effect relationships. “Time” refers to the chronological spans considered when weighing the effects of events and cause-and-effect relationships. It may be that because experts have a broader sense of scale and time, they pay more attention to background information.

We hope that in the future the methodology used in this study can contribute to the development of more effective learning methods for novices in business.

ACKNOWLEDGMENT

“Website Flaming!” used in this study was a fiscal 2009–2010 Chuo University Joint Research project, funded by the Chuo University Joint Research Grant for “advanced case study research in education for the acquisition and exchange of practical knowledge.” Contributors to the creation of “Website Flaming!” were Assistant Professor Mikako Ogawa of the Tokyo University of Marine Science and Technology, Director Akihiko Yanagizaka, and Manga Artist Haruki Ogawa.

This work was supported by JSPS KAKENHI Grant Numbers 26750088, 23501059, 25240048.

REFERENCES

- [1] M. Easterby-Smith, R. Snell, and S. Gherardi, “Organizational Learning: Diverging Communities of Practice?,” *Management Learning*, vol. 29, pp. 259-272, 1998.
- [2] M. Matsuo and T. Kusumi, “Salesperson’s Procedural Knowledge, Experience and Performance: An Empirical Study in Japan,” *European Journal of Marketing*, vol. 36, pp. 840-854, 2002.
- [3] T. Kato and T. Fukuda, “Visual search strategies of baseball batters: eye movements during the preparatory phase of batting,” *Perceptual and Motor Skills*, vol. 94, pp. 380-386, 2002.
- [4] M. T. H. Chi, R. Glaser, and E. Rees, “Expertise in problem solving,” *Advances in the psychology of human intelligence*, vol. 1, R.J. Sternberg, Ed. Hillsdale, NJ: Erlbaum, pp. 7-75, 1982.
- [5] A. Orita, A. Yoshikawa, and T. Terano, “Conducting Situated Intelligence Training ? Practices of executive training using manga,” *Proceedings of International Conference Rethinking Business and Business Education*, Chois, Greece, University of Aegean, 2011. (CD-ROM)
- [6] A. Orita, A. Yoshikawa, and T. Terano, “Practical IS-Education using manga,” *OASIS2011 IFIP8.2 Pre-ICIS Workshop*, Shanghai, China, 2011
- [7] A. Yoshikawa, A. Orita, and T. Terano, “Design of Situated Intelligence Training - A method for executive training using manga,” *Proceedings of International Conference Rethinking Business and Business Education*, Chois, Greece, University of Aegean, 2011. (CD-ROM)
- [8] A. Orita, A. Yoshikawa, and T. Terano, “Workshop on Business Csse Studies Using Narrative Approach With Manga Text,” *KMO2011 Sixth International KMO Conference, Knowledge Management in Organizations*, 2011.(workshop)
- [9] K. A. Ericsson and A. C. Lehmann, ”Expert and exceptional performance: Evidence of maximal adaptation to task constraints,” *Annual review of psychology*, vol.47, pp. 273-305, 1996.
- [10] J. Misumi and M. F. Peterson, “The Performance-Maintenance (PM) Theory of Leadership: Review of a Japanese Research Program,” *Administrative Science Quarterly*, vol. 30, pp. 198-223 , 1985.
- [11] J. Misumi, “トップマネジメントリーダーシップのPMスケール作成とその妥当性の研究 (Development and validation of PM scale of top management leadership),” *Hakutoshobo, Shoshikikagaku*, vol. 20, pp 91-104, 1987. (in Japanese)

TABLE II. CATEGORIZATION OF NOVICE INTERVIEW RESPONSES

		Scale			
		Tanaka's Attitude and Nature	Tanaka's Relationship with Staff	Tanaka's Relationship with the Outside World	Overall Organization
Time	The Past				
	Within the Story	<ul style="list-style-type: none"> •I can imagine what he is thinking. •He didn't do it early enough. •He should move earlier. Etc. 	<ul style="list-style-type: none"> •Pushes the ultimate responsibility onto his staff. •Runs meetings efficiently. •He gives his management opinion and he shares it. Etc. 	<ul style="list-style-type: none"> •He's good at external PR. •He knows how to talk to people outside the company. •He's a vigorous salesman. Etc. 	<ul style="list-style-type: none"> • Good benefits.
	The Future				

TABLE III. CATEGORIZATION OF EXPERT INTERVIEW RESPONSES

		Scale			
		Tanaka's Attitude and Nature	Tanaka's Relationship with Staff	Tanaka's Relationship with the Outside World	Overall Organization
Time	The Past		<ul style="list-style-type: none"> •They didn't prepare for the meeting by working out an integrated approach internally beforehand. 		<ul style="list-style-type: none"> •The server should have been reinforced. •If it's known beforehand that usage is going to suddenly increase, shouldn't they devise some way like using filters so that, for at least that day, the servers are down as little as possible? •They're not going to let the servers go down, but the whole system is broken.
	Within the Story	<ul style="list-style-type: none"> •He is not up to the job. •He tries to handle difficult situations alone. •On the positive side, it's good that he tries to take responsibility for problems. Etc. 	<ul style="list-style-type: none"> •He proactively adopts proposals from staff. •He communicates with staff over yakiniku dinners. •He is surprisingly well liked by his staff. Etc. 	<ul style="list-style-type: none"> •Shouldn't he be clearly questioning the business mode!? They would be overlooking something, because they don't have B to C experience. He goes around impressing them with his strong points. •He is concerned about social issues. •His excuses are long-winded! Etc. 	<ul style="list-style-type: none"> •Awareness of the company's financial struggles is not shared. •Considering the revenue, expenses are too high (the building, conference rooms). •Staff members are not on the same page. Etc.
	The Future	<ul style="list-style-type: none"> •He's thinking about new businesses 		<ul style="list-style-type: none"> •If there really were problems, to respond that way is a bad idea. 	

Compound Noun Polysemy and Sense Enumeration in WordNet

Abed Alhakim Freihat

Dept. of Information Engineering
and Computer Science
University of Trento,
Trento, Italy

Email: fraihat@disi.unitn.it

Biswanath Dutta

Documentation Research
and Training Centre
Indian Statistical Institute (ISI)
Bangalore, India

Email: bisu@drtc.isibang.ac.in

Fausto Giunchiglia

Dept. of Information Engineering
and Computer Science
University of Trento,
Trento, Italy

Email: fausto@disi.unitn.it

Abstract—Sense enumeration in WordNet is one of the main reasons behind the problem of high polysemous nature of WordNet. The sense enumeration refers to misconstruction that results in wrong assigning of a synset to a term. In this paper, we propose a novel approach to discover and solve the problem of sense enumerations in compound noun polysemy in WordNet. The proposed solution reduces the number of sense enumerations in WordNet and thus its high polysemous nature without affecting its efficiency as a lexical resource for natural language processing.

Keywords—Polysemy; wordNet; compound nouns; sense enumeration.

I. INTRODUCTION

WordNet or Princeton WordNet [1] is a machine readable online lexical database for the English language. Based on psycholinguistic principles, WordNet has been developed since 1985 by linguists and psycholinguists as a conceptual dictionary rather than an alphabetic one [2].

Compound nouns are multi-words or collocations that consist of modified nouns and noun modifiers. One such example is the noun *nerve center*, where the *center* is the modified noun and *nerve* is the noun modifier. Compound noun polysemy in WordNet refers to the cases where we use the modified noun to refer to several different compound nouns such as using the modified noun *center* to refer to *nerve center* and *shopping center* [3]. The meanings of a compound noun polysemous term may correspond to a specialization polysemy [4] [5], metonymy [6] [7], or they are just sense enumerations, i.e., a misconstruction that results in wrong assignment of a synset to a term. Assigning the term *center* to the following two synsets is an example of sense enumerations:

- #3 *center, nerve center*: a cluster of nerve cells governing a specific bodily process.
- #15 *plaza, mall, center, shopping mall, shopping center*: mercantile establishment consisting of a carefully landscaped complex of shops...

The problem of sense enumerations in compound nouns is that they are a source of noise rather than a source of knowledge when using WordNet as a source for natural language processing (NLP) and knowledge-based applications, especially Information Retrieval (IR) [8] and semantic search [9].

Although specific instances of the compound noun polysemy have been addressed when solving the problem of specialization polysemy [4] [5] [10] or metonymy [7] [11] [12] [13], no or little research has been made towards solving the problem of compound noun polysemy as a problem of sense enumeration in WordNet.

In this paper, we discuss the problem of sense enumerations in compound noun polysemy in general and propose a semi-automatic method which allows us to discover and resolve sense enumerations in compound noun polysemy. The proposed solution is a cleaning process that reduces the number of sense enumerations in WordNet and thus its high polysemous nature without affecting its efficiency as a lexical resource.

The paper is organized as follows. In Section II, we discuss the compound noun polysemy and the relation between this kind of polysemy and the high polysemous nature in WordNet. In Section III, we briefly discuss the state of the art. In Section IV, we introduce the formal definitions that we used in our approach. In Section V, we present a semi automatic method for solving the problem of sense enumerations in WordNet in the case of compound noun polysemy. In Section VI, we discuss the results of the proposed method. In Section VII, we conclude the paper.

II. SENSE ENUMERATIONS IN COMPOUND NOUN POLYSEMY

A term in wordNet can be a single word such as *center* or a collocation such as *nerve center*. In the case of nouns, collocations correspond to compound nouns. A compound noun contains two parts.

- 1) *noun adjunct/modifier*: a noun that modifies another noun in a compound noun.
- 2) *noun head/modified noun*: the modified noun in a compound noun.

For example, the noun *head* is the noun adjunct and *word* is the modified noun in the compound noun *head word*. WordNet contains 104290 nouns. These nouns belong to 74314 synsets. The number of compound nouns is 58946 and the number of the synsets that contain at least one compound noun is 40560. That means, more than 56% of the nouns in WordNet are compound nouns and more than 45.4% of the synsets contain compound nouns. *Compound noun polysemy* refers to the cases, where we use the modified noun to refer to several different compound nouns. The number of the compound

polysemous nouns in WordNet is 3407. These nouns belong to 4918 polysemous synsets. Compound noun polysemy in WordNet belong to the following three groups:

- 1) *Metonymy*: Corresponds to the metonymy polysemy cases where the modified noun belongs to two synsets, one of these synsets is base meaning and the other is derived meaning. For example, the compound noun polysemy between the following two synsets is an instance of metonymy.

#2 cherry, cherry tree: any of numerous trees and shrubs producing a small fleshy round fruit with a single hard stone.

#3 cherry: a red fruit with a single hard stone.

- 2) *Specialization Polysemy*: Corresponds to the specialization polysemy cases where the modified noun belongs to two synsets, one of these synsets is a more general meaning of the other or both synsets are more specific meanings of a third synset. For example, the compound noun polysemy between the following two synsets is an instance of specialization polysemy.

#1 red laver, laver: edible red seaweeds.

#2 sea lettuce, laver: seaweed with edible translucent crinkly ...

- 3) *Sense enumeration*: Sense enumeration means a misconstruction that results in wrong assignment of a synset to a term, i.e., assignment the noun modifier or the modified noun as a synonym of the compound noun itself. For example, assigning the the term head as a synonym to the compound nouns in the following synsets is an instance of sense enumerations.

#8 fountainhead, headspring, head: the source of water from which a stream arise.

#9 head, head word: *grammar* the word in a grammatical constituent that plays the same grammatical role as the whole constituent.

#13 principal, school principal, head teacher, head: the educator who has executive authority for a school.

#16 promontory, headland, head, foreland: a natural elevation (especially a rocky one that juts out into the sea).

#21 headway, head: forward movement.

#27 read/write head, head: a tiny electromagnetic coil and metal pole used to write and read magnetic patterns on a disk.

#32 drumhead, head: a membrane that is stretched taut over a drum.

In general, using the modified noun to refer to the compound noun itself is usual in natural language. In such cases, we use the context to understand and disambiguate the modified noun. An important question here is the relation between the lexicon and the ability to understand and disambiguate

the modified noun. The issue is whether compound nouns and their corresponding modified nouns should be stored as synonyms in the lexicon. In natural language processing, do we need a lexical database that assigns each modified noun as a synonym to its corresponding compound nouns to be able to disambiguate the cases in which we use modified nouns to refer to compound nouns?

If we need to explicitly store the synonymity between each modified noun and its corresponding compound nouns, then the polysemous nouns in WordNet should be at least 56% and the polysemous synsets at least 45% just to store this information. For example WordNet contains 135 non polysemous synsets in which the term head is a noun modifier or modified noun of a compound noun. That means, the number of the senses of the term head in WordNet should be 168 (head has 33 senses in WordNet). For example the term head should be synonymous to the terms department head, head of household, head of state, head nurse, human head, nominal head, hammerhead, axe head, spearhead, magnetic head, ...

In this approach, we argue that using a noun adjunct/modified noun to refer to its corresponding compound noun is similar to the use of anaphoric pronouns [14] (he, she, it, ...). This means that the disambiguation of polysemous modified nouns depends on the context rather on the used lexicon. In this sense, we may call a noun adjunct/modified noun that refers to a compound noun an *anaphoric term*. Anaphoric pronouns and anaphoric terms are similar in the following aspects:

- 1) Anaphoric pronouns and anaphoric terms are usually used to avoid repetition of the same word.
- 2) Anaphoric pronouns and anaphoric terms are usually ambiguous.
- 3) Using and understanding of anaphoric pronouns and anaphoric terms depends on a term that precedes them.
- 4) Anaphoric pronouns and anaphoric terms usually need a disambiguation process which allows to bind them to their corresponding referred term in the discourse.

In point 3, the discourse dependency of anaphoric terms means that an anaphoric term is used to refer to another (explicit or implicit) term in the context that enables disambiguating the reference term. That means, without (the explicit or implicit) referred term, the anaphoric term has no meaning or its meaning can not be disambiguated. We think that the referred term is the nearest understood compound noun. Thus, using and understanding the reference term is dependent on a compound noun that can be understood from the discourse and does not depend on storing the polysemy relation between the referred term and the the reference term in the lexicon.

Similar to anaphoric pronouns in point 4, anaphoric terms need to be disambiguated. Anaphoric pronoun disambiguation is called *anaphoric resolution* which is a syntactic process that binds the pronouns to their corresponding referred terms.

Our hypothesis in this approach is that reference term disambiguation is similar to pronoun disambiguation. That means, removing the anaphoric terms from WordNet in all compound noun polysemy cases reduces the sense enumerations without affecting its efficiency as a lexical resource for NLP tools.

III. RELATED WORK

In general, the polysemy approaches address the Compound noun polysemy as a sub case of metonymy and specialization polysemy. These approaches did not address solving the sub cases of compound noun polysemy that correspond to sense enumerations. In the following, we summarize the most prominent polysemy approaches for solving metonymy and specialization polysemy.

CORELEX [11] is a database of systematic polysemy classes (based on the generative lexicon theory [15]). These classes are combinations of 39 basic types that reside at the top level of WordNet hierarchy such as {animal, plant, food, attribute, state, artifact, ...}. The idea is that metonymy cases can be underspecified to one of these classes. Systematic polysemous meanings are systematic and predictable. The polysemy type of the term *banana* in the following example is systematic since the meaning *food* can be predicted from the *plant* meaning and so these two meanings of *banana* belong to the systematic class *plant#food*.

- #1 *banana, banana tree*: any of several tropical and subtropical treelike herbs of the genus *Musa* having a terminal crown of large entire....
- #2 *elongated crescent-shaped yellow fruit with soft sweet flesh*.

The semantic relations extraction approaches are regular polysemy [16] approaches that attempt to extract implicit semantic relations between the polysemous senses via regular structural patterns. The basic idea in these approaches is that the implicit relatedness between the polysemous terms corresponds to variety of semantic relations. Extracting these relations and making them explicitly should improve wordNet [12]. These approaches refine and extend CORELEX patterns to extract the semantic relations. Beside the structural regularity, these approaches exploit also the synset gloss [4] and the cousin relationship [7] [12] in WordNet. For example, the approach described in [4] exploits synset glosses to extract auto-referent candidates. The approach described in [7] uses several rules, such as *ontological bridging* [7] to detect relations between the sense pairs.

In general, the extracted relations in the semantic relations extraction approaches are similar. For example, we find the relations *similar to* or *color of* in the results of the approach in [4]. The results in [7] contains relations such as *contained in*, *obtain from*. Similarly, the result in [12] contain relations such as *fruit of*, *tree of*.

Specialization polysemy approaches such as [3] [4] are regular polysemy approaches that attempt to transform the implicit hierarchical relation between the synsets from lexical level to the semantic level. The approach described in [10] [5] considers representing the hierarchical relation at lexical level as a kind of sense enumeration that leads to high polysemy and information lost. An example for transforming the hierarchical relation from lexical level to the semantic level is shown in Figure 1.

IV. APPROACH DEFINITIONS

In this section, we present the definitions that we use in our approach. We start with the basic definitions. We define terms as follows.

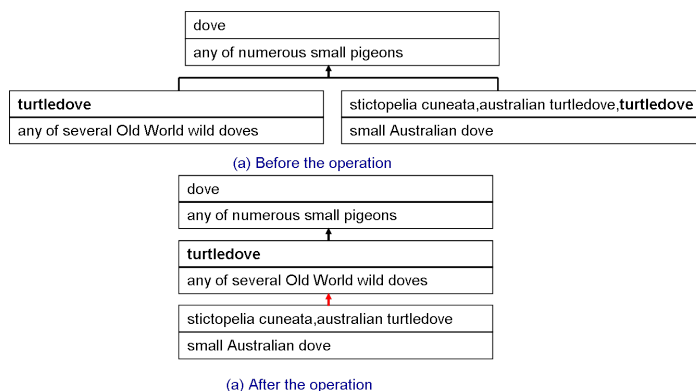


Figure 1. Example of transforming the hierarchical relation from the lexical level to the semantic level

Definition 1: (Term) A term T is a triple $\langle \text{Lemma}, \text{Cat}, \text{T-Rank} \rangle$, where

- Lemma is the term lemma, i.e., the orthographic string representation of the term;
- $\text{Cat} \in \{\text{noun, verb, adjective, adverb}\}$ is the grammatical category of the term;
- T-Rank is the term rank, i.e., a natural number >0 .

T-Rank is used to reflect which is the preferred term of a synset. For example, *man* and *adult male* in the following synset correspond to the following term instances: $\langle \text{Lemma: "man"}, \text{Cat: noun, T-Rank: 1} \rangle$ and $\langle \text{Lemma: "adult male"}, \text{Cat: noun, T-Rank: 2} \rangle$.

- #1 *man, adult male*: an adult person who is male (as opposed to a woman).

In the following, we define wordNet synsets.

Definition 2: (WordNet synset) A synset S is defined as $\langle \text{Cat}, \text{Terms}, \text{Label}, \text{Gloss} \rangle$, where

- $\text{Cat} \in \{\text{noun, verb, adjective, adverb}\}$ is the grammatical category of the synset ;
- Terms is an ordered list of synonymous terms that have the same grammatical category as the synset grammatical category;
- $\text{Label} \in \text{Terms}$ is the preferred term of the synset, i.e., the term whose T-Rank = 1;
- Gloss is a natural language text that describes the synset.

A term is polysemous if it is found in the terms of more than one synset. We define polysemous term as follows.

Definition 3: (polysemous term) A term $t = \langle \text{Lemma}, \text{Cat}, \text{T-Rank} \rangle$ is polysemous if there is a term t' and two synsets s and s' , $s \neq s'$ such that

- $t \in s.$ Terms and $t' \in s'.$ Terms
- $t.$ Lemma = $t'.$ Lemma
- $t.$ Cat = $t'.$ Cat.

A synset is polysemous if it contains at least one polysemous term. We define polysemous synsets as follows.

Definition 4: (polysemous synset) A synset s is polysemous if one of its terms is a polysemous term.

It is possible for two polysemous synsets to share more than one term. Two polysemous synsets and their shared terms constitute a polysemy instance. In the following, we define polysemy instances.

Definition 5: (polysemy instance) A polysemy instance is a triple $[\{T\}, s_1, s_2]$, where s_1, s_2 are two polysemous synsets that have the terms $\{T\}$ in common.

The second step is to formalize the case where we have a polysemy instance of a compound noun.

Definition 6: (compound noun polysemous term) A term t is compound noun polysemous term of a term t' if t is the noun adjunct or the modified noun of t' .

For example, the term *center* is a compound noun polysemous term of the term *nerve center*. In the following, we define a compound noun polysemous synset.

Definition 7: (compound noun polysemous synset) A synset s is compound noun polysemous if it contains a compound noun polysemous term.

For example, the following synset is a compound noun polysemous synset.

- #7 center, centre, nerve center, nerve centre: a cluster of nerve cells governing a specific bodily process

In the following, we define compound noun polysemy instance.

Definition 8: (compound noun polysemy instance) A polysemy instance $I = [\{T\}, s_1, s_2]$ is compound noun polysemy instance if s_1 or s_2 is a compound noun polysemous synset.

For example, $[\{center, centre\}, \#7, \#8]$ is a compound noun polysemy instance because #7 is a compound noun polysemous synset.

- #7 center, centre, nerve center, nerve centre: a cluster of nerve cells governing a specific bodily process
- #8 center: the middle of a military or naval formation

The third step is to define the structural patterns which allow us to identify specialization polysemy instances in compound nouns.

Definition 9: (structural pattern) A structural pattern of a polysemy instance $I = [\{T\}, s_1, s_2]$ is a triple $P = \langle r, p_1, p_2 \rangle$, where

- a) r is the least common subsumer of s_1 and s_2 ;
- b) p_1 and p_2 are children of r .
- c) p_1 subsumes s_1 and p_2 subsumes s_2

For example, $\langle mercantile\ establishment, marketplace, shop \rangle$ is the structural pattern of the polysemy instance $[\{bazaar; bazar\}, s_1, s_2]$ as shown in Figure 2.

The following definition allows us to define the specialization polysemy instances in compound nouns.

Definition 10: (Specialization Polysemy instance) A compound noun polysemy instance $I = [\{T\}, s_1, s_2]$ is a specialization polysemy instance if its structural pattern $p = \langle r, p_1, p_2 \rangle$ has one of the following forms $\langle r, s_1, s_2 \rangle$, $\langle r, s_1, p_2 \rangle$ or $\langle r, p_1, s_2 \rangle$ as illustrated in Figure 3.

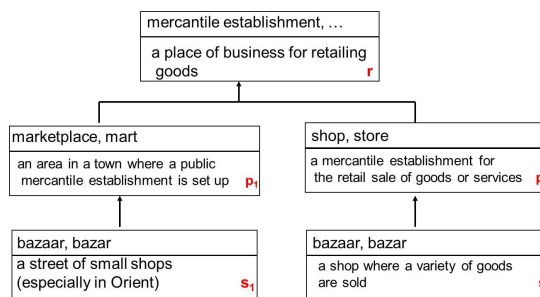


Figure 2. Example of a structural pattern

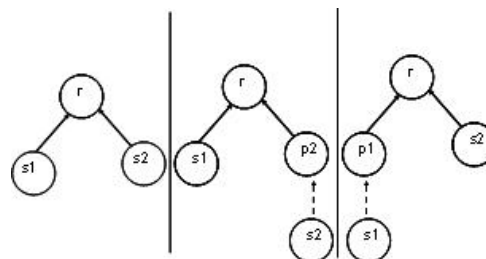


Figure 3. Specialization polysemy pattern

In the following, we define compound noun polysemy instances that belongs to metonymy by using CORELEX structural patterns.

Definition 11: (CORELEX structural pattern) CORELEX structural pattern is a sequence of synset labels separated by # where each synset label corresponds to a synset in WordNet.

For example, a CORELEX structural pattern is *plant#fruit*. In the following, we define CORELEX polysemy classes.

Definition 12: (CORELEX polysemy class) Let $p = p_1 \# p_2$ be a CORELEX pattern. The polysemy class of p is defined as the set of all polysemy instances $\{I = [\{T\}, s_1, s_2] \mid s_1 \text{ is subsumed by } p_1 \text{ and } s_2 \text{ is subsumed by } p_2\}$

For example, the polysemy instance $\{I = [\{peach\}, \#1, \#3]\}$ belongs to the polysemy class of CORELEX structural pattern *plant#fruit* because the synset #1 is subsumed by *plant* and #3 is subsumed by *fruit*.

- #1 peach, peach tree, Prunus persica: cultivated in temperate regions.
- #3 peach: downy juicy fruit with sweet yellowish or whitish flesh.

In the following, we define the notion of metonymy instance.

Definition 13: (Metonymy instance) A polysemy instance I is a metonymy instance if it belongs to some CORELEX polysemy class.

Finally, we define sense enumeration in compound noun polysemy.

Definition 14: (Sense enumeration in compound noun polysemy) A compound noun polysemous term in a compound noun polysemous synset s_1 is considered to be a sense enumeration if the following hold:

- a) s_1 is a compound noun polysemous synset;

- b) There is no polysemy instance $I = [\{T\}, s_1, s_2]$ such that I is a metonymy or a specialization polysemy instance.

V. DISCOVERY AND ELIMINATION OF SENSE ENUMERATIONS IN COMPOUND NOUNS

In this section, we describe the discovery and elimination of sense enumerations in compound nouns. This is performed by a semi-automatic process that includes the following steps.

P1 Discovery of sense enumerations in Compound Nouns: Sense enumerations discovery in compound nouns is performed semi-automatically as follows.

- 1) **Sense enumeration candidates discovery:** This step is automatic and performed by deploying an algorithm that returns sense enumeration candidates in compound noun polysemous nouns.
- 2) **Excluding of false positives:** This step is manual where we exclude the false positives from the output of the algorithm in the previous step. For example, we exclude term abbreviations.

P2 Elimination of sense enumerations: This step is automatic and performed by deploying an algorithm which allows us to eliminate sense enumerations in the identified cases by removing the polysemous noun modifier and keeping the compound noun.

A. Discovery of sense Enumerations in Compound Nouns

In the following, we discuss the algorithm that we deployed in the discovery of sense enumerations in compound nouns. The algorithm returns a hash map of compound noun polysemous terms and senses enumeration candidates according to definition 14 and it works as follows:

- 1 It retrieves all compound noun polysemous terms in WordNet.
- 2 It iterates over all retrieved compound nouns to identify sense enumeration candidates as follows. For each retrieved compound noun term:
 - 2.a It computes a list of the term synsets.
 - 2.b It computes a list of the polysemy instances of each of the retrieved synsets.
 - 2.c It checks if any of the polysemy instances of the synset is a specialization polysemy instance according to definition 10 or a metonymy instance according to definition 13.
 - 2.d if none of the polysemy instances of the synset is specialization polysemy or metonymy, the synset is considered as a sense enumeration according to definition 14 and added to the sense enumeration list of the term.
- 2.e The compound noun polysemous term and its corresponding sense enumerations are stored in a hash map.
- 3 The algorithm returns the hash map that corresponds to the compound noun polysemous terms and their corresponding sense enumerations.

B. Excluding of False Positives

The input of this phase is the output of the algorithm `senseEnumerationsDiscovery`. The task of this phase is to exclude false positives. Experimentally, it turns out that the false positives can be classified into the following two groups:

- 1) **Missing adjunct noun/modified noun synset:** In some cases, a synset of the adjunct noun or the modified noun is missing. Such cases are excluded. For example, none of the 6 synsets of the term `party` can be considered as a general meaning of the term `political party` in the following synset.

```
# party, political party: an
  organization to gain political
  power.
```

- 2) **Term abbreviations:** Since the algorithm in the previous step uses the string function to test compound noun polysemy, the algorithm returns polysemy instances that include term abbreviations as compound noun polysemy instances. For example, the term `mil` is abbreviation of the terms `milliliter` and `millilitre` in the following synset.

```
# milliliter, mil, ml, cubic
  centimeter, cc: a metric unit
  of volume equal....
```

C. Elimination of Sense Enumerations in Compound Nouns

In this step, we eliminate the sense enumerations by removing the polysemous modified nouns. For example, the result of applying the function on `head` and the synset #32 is the synset #32':

```
#32 drumhead, head: a membrane that is
  stretched taut over a drum.
#32' drumhead: a membrane that is stretched
  taut over a drum.
```

VI. RESULTS AND EVALUATION

In the following, we present the results of our approach. Table I shows the results of the compound noun polysemy discovery algorithm that returned 2270 possible compound noun polysemous terms. These terms belong to 2952 synsets. The total number of compound noun polysemous instances is 11650 instance. Table II shows the results of the man-

TABLE I. RESULTS OF THE DISCOVERY ALGORITHM

#Compound noun polysemous terms	2270
#Compound noun polysemous synsets	2952
#Compound noun polysemous instances	11650

ual validation process, where the synsets of 1905 terms are classified to be sense enumerations. These terms belong to 2547 synsets. These synsets belong to 11088 compound noun polysemy instances. In Table III, we give an overview about

TABLE II. MANUAL VALIDATION RESULTS

#Compound noun polysemous terms	1905
#Compound noun polysemous synsets	2547
#Compound noun polysemous instances	11088

number of nouns, noun senses and noun synsets in resulting

WordNet after applying the disambiguation algorithm on the nouns in the WordNet 2.1. The table shows the reduction of

TABLE III. DISAMBIGUATION ALGORITHM RESULTS

	#Nouns	#Synsets	#Senses
Before Applying the Algorithm	104290	74314	130207
After Applying the Algorithm	104290	74314	127660

WordNet senses from 130207 to 127660. The average sense per noun before applying our algorithm is 1.25. Applying our algorithm reduces sense number per noun to 1.22.

A. Evaluation

To evaluate our approach, 200 synsets have been evaluated by two evaluators. In Table IV, we report the statistics of the evaluation, where we show the following:

- Total agreement*: Measures the number of polysemy instances where both evaluators agrees with our approach (corresponds to second row in the table).
- Partial agreement* Measures the number of polysemy instances where the at least one of the evaluators agrees with our approach (corresponds to third and fourth rows in the table).
- Disagreement* Measures disagreement between the approach and the evaluators (corresponds to last row in the table).

In Table IV, a refers to our approach, e_1, e_2 refer to evaluator1 and evaluator 2 respectively.

TABLE IV. EVALUATION RESULTS

Evaluators Agreement	Result
$a = e_1 \wedge a = e_2$	161 (80.5%)
$a = e_1$	172 (86%)
$a = e_2$	177 (88.5%)
$a \neq e_1 \wedge a \neq e_2$	12 (6%)

VII. CONCLUSION AND FUTURE WORK

In this paper, we have introduced a new approach for solving the problem of sense enumerations in compound noun polysemy, where we have removed the sense enumerations in compound nouns in WordNet and thus reduced the high polysemy in compound nouns. The proposed solution is a necessary step that should be included in any approach for solving the polysemy problem in WordNet because the sense enumerations in compound nouns is a source of noise rather than a source of knowledge that affects the quality of WordNet as a source for NLP and knowledge-based applications.

Although the manual treatment in the approach guarantees the quality of the approach, we plan to run an indirect evaluation to test the effects of our approach in terms of precision and recall as a future work. As future work, we plan also to examine the relation between sense enumeration and missing terms in WordNet especially when a synset contains a modified noun and the compound noun itself is missing in the synset. For example, solving the sense enumeration problem in the following two meanings of the term `head`, we add the missing terms `bony pelvis` and `head of muscle` in the following two synsets respectively.

- #25 `head`:the rounded end of a bone that bits into a rounded cavity in another bone to form a joint.
- #26 `head`: that part of a skeletal muscle that is away from the bone that it moves.

Acknowledgment. The research leading to these results has received partially funding from the European Community's Seventh Framework Program under grant agreement n. 600854, Smart Society (<http://www.smart-society-project.eu/>).

REFERENCES

- G. A. Miller, "Wordnet: A lexical database for english," Commun. ACM, vol. 38, no. 11, Nov. 1995, pp. 39–41. [Online]. Available: <http://doi.acm.org/10.1145/219717.219748>
- G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to WordNet: an on-line lexical database," International Journal of Lexicography, vol. 3, no. 4, 1990, pp. 235–244. [Online]. Available: <http://wordnetcode.princeton.edu/5papers.pdf>
- A. A. Freihat, "An organizational approach to the polysemy problem in wordnet," PhD thesis, University of Trento, 2014.
- L. Barque and F.-R. Chaumartin, "Regular polysemy in wordnet," JLCL, vol. 24, no. 2, 2009, pp. 5–18. [Online]. Available: <http://dblp.uni-trier.de/db/journals/ldvi/ldvf24.html#BarqueC09>
- A. A. Freihat, F. Giunchiglia, and B. Dutta, "Solving specialization polysemy in wordnet," International Journal of Computational Linguistics and Applications, vol. 4, no. 1, jan-june 2013.
- W. Peters and I. Peters, "Lexicalised systematic polysemy in wordnet," in LREC. European Language Resources Association, 2000. [Online]. Available: <http://dblp.uni-trier.de/db/conf/lrec/lrec2000.html#PetersP00>
- T. Veale, "A non-distributional approach to polysemy detection in wordnet," [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/download?rep=rep1&type=pdf&doi=10.1.1.146.5566>
- R. Mandala, T. Tokunaga, and H. Tanaka, "Complementing wordnet with roget's and corpus-based thesauri for information retrieval," in EACL. The Association for Computer Linguistics, 1999, pp. 94–101. [Online]. Available: <http://dblp.uni-trier.de/db/conf/eacl/eacl1999.html#MandalaTT99>
- F. Giunchiglia, U. Kharkevich, and I. Zahirayeu, "Concept search: Semantics enabled syntactic search," in Proceedings of the Workshop on Semantic Search (SemSearch 2008) at the 5th European Semantic Web Conference (ESWC 2008), June 2, 2008, Tenerife, Spain, ser. CEUR Workshop Proceedings, S. Bloehdorn, M. Grobelnik, P. Mika, and D. T. Tran, Eds., vol. 334. CEUR-WS.org, 2008. [Online]. Available: <http://ceur-ws.org/Vol-334/paper-10.pdf>
- A. A. Freihat, F. Giunchiglia, and B. Dutta, "Regular polysemy in wordnet and pattern based approach," International Journal On Advances in Intelligent Systems, no. 3&4, jan 2013.
- P. Buitelaar, "Corelex: Systematic polysemy and underspecification," PhD thesis, Brandeis University, Department of Computer Science, 1998.
- P. W., "Detection and characterization of figurative language use in wordnet," PhD thesis, Natural Language Processing Group, Department of Computer Science, University of Sheffield, 2004.
- S. N. Kim and T. Baldwin, "Word sense and semantic relations in noun compounds," ACM Trans. Speech Lang. Process., vol. 10, no. 3, Jul. 2013, pp. 9:1–9:17. [Online]. Available: <http://doi.acm.org/10.1145/2483969.2483971>
- J. Zheng, W. W. Chapman, R. S. Crowley, and G. K. Savova, "Coreference resolution: A review of general methodologies and applications in the clinical domain," Journal of Biomedical Informatics, vol. 44, no. 6, 2011, pp. 1113 – 1122. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S153204641100133X>
- J. Pustejovsky, "The generative lexicon," Computational Linguistics, vol. 17, 1991.
- A. J., "Regular polysemy," Linguistics, 1974, pp. 5–32.

Automatic Diagrammatic Multiple Choice Question Generation from Knowledge Bases

Khalil Bouzekri, Liu Qiang, Husam N. Yasin, Benjamin Chu Min Xian, Dickson Lukose

Artificial Intelligence Department, MIMOS Berhad, Kuala Lumpur, Malaysia

khalil.ben@mimos.my, qiang.liu@mimos.my, husam.yasin@mimos.my, mx.chu@mimos.my, dickson.lukose@mimos.my

Abstract—In this paper, we present a methodology to generate diagrammatic multiple choice questions from a knowledge base. When considering a knowledge base, the main strategies discussed in the literature are class-based, property-based and terminology-based, and the generated multiple choice questions are typical with all the choices (correct answer and distractors) being of atomic types such as plain text. In this paper, we introduce graph-based strategies, enabling the generation of choices (correct answer and distractors) in the form of complex structures such as diagrams, and discuss different approaches to take into account difficulty levels when generating the questions.

Keywords—Question Generation; MCQ; Diagram; Graph; Question Difficulty; Knowledge Base; Ontology; Islamic Finance.

I. INTRODUCTION

A Multiple Choice Question (MCQ) comprises of a short text describing the question, and a number of alternative choices as answer, where usually one of the choices is the correct answer and the others are wrong alternative choices called distractors. MCQs are popular in e-learning systems due to several characteristics: simplicity of generation (header and choices), ease for scalability (systems can generate a large number of different MCQs), and automation and objectivity of the assessments.

The use of the semantic web to generate questions has been studied at length [1][2][3][8][9]. Indeed, domain ontologies can be considered a proper formalism as the basis for automatic generation of MCQs. Ontology contains domain knowledge in the form of concepts, instances and relationships between concepts and/or instances. Moreover, the concepts and relationships are structured in a hierarchical manner based on their semantics. Papasalouros et al. proposed an ontology-based approach to automatically generate MCQs [1]. They used Natural Language Processing (NLP) techniques to generate the question header, and a series of strategies based on the structure of the ontology (class-based, property-based and terminology-based) to generate the correct answer and the distractors. Their strategies were further enriched and implemented as a plugin in Protégé [2]. Cubric and Tomic extended the work by considering new elements such as annotations, and leveraging on question templates instead of NLP to generate the question header [3].

However, all the research has been focusing on the generation of “simple” MCQs, which means MCQs containing a question header and choices (correct answer and distractors) in an atomic format (e.g., plain text, numbers). To the best of our knowledge, more complex structures as choices (e.g., diagrams) have not been studied when automatically generating MCQs. Therefore, the main novelty of our research work resides in two elements: firstly, we propose a more complex type of MCQ that we call “Diagrammatic MCQ”, in which the correct answer is a graph composed of labelled nodes and arcs (e.g., a graph representing a diagram) generated from a knowledge base, and the distractors are generated by combining existing strategies from the literature and new strategies relying on the structure of the graph itself; secondly, we discuss the difficulty levels of the different question generation strategies.

The paper is structured as follows: Section 2 recalls basic notions utilized throughout the paper. In Section 3, we introduce the diagrammatic MCQs and discuss the utilized strategies. Section 4 discusses the difficulty of generated questions. Finally, the prospects of this work are outlined in Section 5.

II. BASIC NOTIONS

In this paper, we use the Islamic Finance Knowledge Base (IFKB) from [4], which comprises a total of 2281 RDF triples [7]. An RDF triple consists of three components: subject, predicate and object. The predicate expresses a relationship between the subject and the object. For example, the relationship between “*Deposit*” and “*CashInvestment*” can be represented using the following triple: “*Deposit subClassOf CashInvestment*” where “*Deposit*” is the subject, *subClassOf* is the predicate and *CashInvestment* is the object (this triple can be read as “Deposit is a specialization/kind of CashInvestment”). Islamic Finance concepts are organized in a hierarchy, as shown in Figure 1: the concept “*Bank*” is a subclass of “*Party*” (also read as “*Bank*” is a child of “*Party*”, “*Party*” is a parent of “*Bank*”), and the siblings of “*Bank*” are “*Bank Client*”, “*Partner*” and “*Supplier*”.

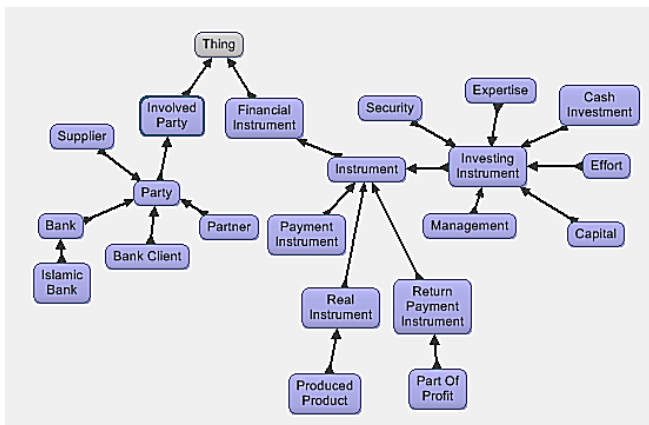


Figure 1. Excerpt of IFKB Concept Hierarchy

IFKB also contains 45 Islamic Finance contract models represented as sets of RDF triples. An Islamic Finance contract model shows the parties involved in the contract, the financial instruments utilized, as well as the process flow. Figure 2 illustrates the Salam contract model in IFKB (rendered using Gruff tool [6]):

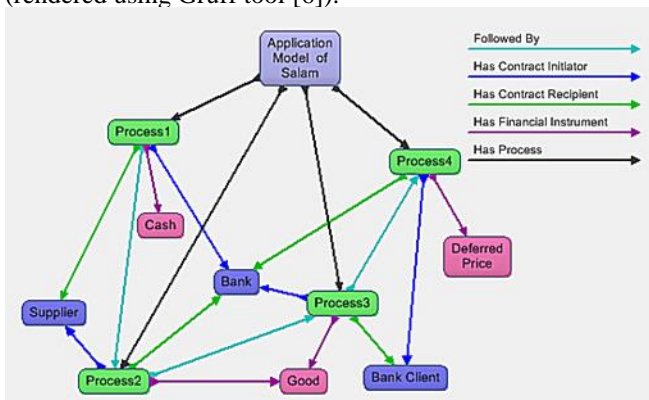


Figure 2. Salam Contract Model in IFKB

There are four processes, where Process1 is followed by Process2, then Process3 and subsequently Process 4. In this application model, the Bank plays the role of a buyer/purchaser as well as seller. The financial instruments used are “Cash”, “Deferred Price” and “Good”, and the involved parties are “Bank Client”, “Bank” and “Supplier”. They are all concepts from IFKB. As an example, a triple is “Process1 HasContractInitiator Bank”. Figure 3 shows the Salam contract model in a diagrammatical form, which is the visualization used to display the DMCQs to the user.

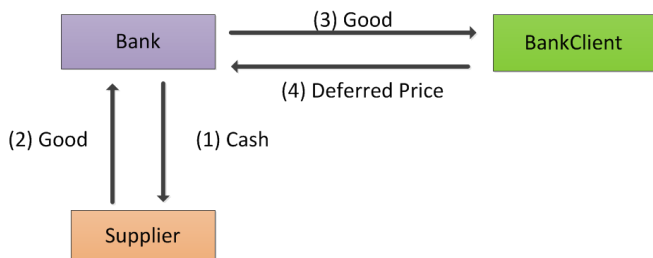


Figure 3. Salam Contract Model in Diagrammatical Form

Salam contract model (lit: forward payment) is the sale of a deferred item in exchange for an immediate (forward) price [5].

III. DIAGRAMMATIC MULTIPLE CHOICE QUESTION GENERATION

Diagrams are useful mechanisms for information summarization and displaying relationships between objects. They are used to explain concepts and amplify understanding. In this section, we present four types of diagrammatic multiple choice question (DMCQ), and discuss the strategies involved to generate these DMCQs such as choosing a question header, choosing a correct answer by extracting a group of concepts and relationships representing a contract model from a knowledge base, and forming the distractors. In the following, we use the Islamic contracts of IFKB to build DMCQs.

A. Diagrammatic Multiple Choice Question

A DMCQ consists of a question header, a correct answer and distractors (wrong answers). In the following, each DMCQ will have one correct answer and three distractors. In DMCQ Type 1, the correct answer is a set of nodes or arcs from the diagram. DMCQ Type 2 is based on conventional fill-in-the-blank questions, where one or more labels, nodes or arcs will be removed from the correct diagram and the user needs to select the correct answer from the given choices to fill in the blank. The correct answer of the DMCQ Type 3 is the original diagram from the domain knowledge base. Finally, the correct answer of DMCQ Type 4 is a partial diagram which is extracted from the original diagram. For each DMCQ type, the distractors are derived from the correct answer using various strategies. The distractors generation strategies and difficulty levels are discussed in Section III. D.

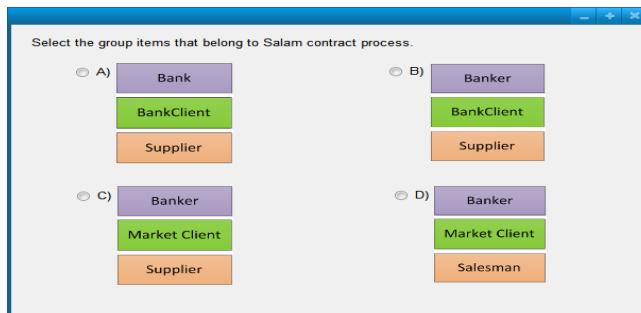


Figure 4. DMCQ Type 1 Sample

Figure 4 shows a sample of DMCQ Type 1, where the question header is “Select the group items that belong to Salam contract process”, and the possible answers are sets of parties. Choice “A” is the correct answer and choices “B”, “C” and “D” are distractors.

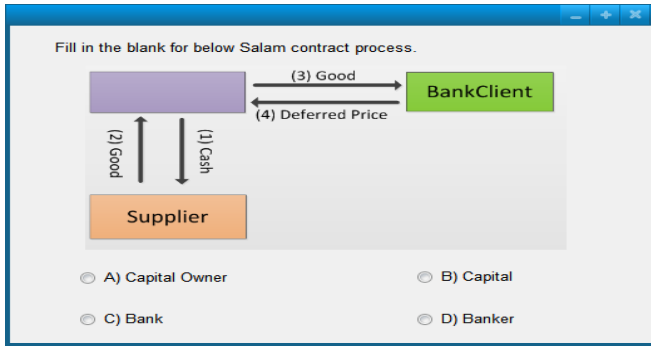


Figure 5. DMCQ Type 2 Sample

A sample of DMCQ Type 2 is shown in Figure 5. The party label “Bank” has been removed from the diagram representing the Salam contract model. In this sample, the modified diagram is part of the question header. Question choice “C” is the correct answer, and “B”, “C” and “D” are distractors.

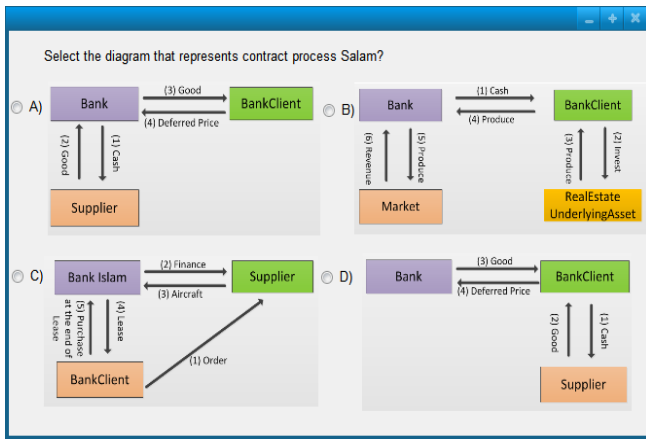


Figure 6. DMCQ Type 3 Sample

Figure 6 shows a sample of DMCQ Type 3, where the correct answer is the diagram of the Salam contract model (choice “A”), and “B”, “C” and “D” are the distractors.

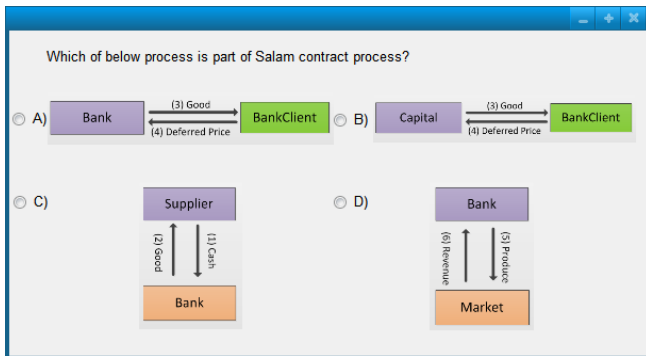


Figure 7. DMCQ Type 4 Sample

DMCQ Type 4 question sample is shown in Figure 7. The correct answer is a subset of the full diagram of Salam

contract model (choice “A”), and choices “B”, “C” and “D” are distractors.

B. Question Header Generation

As discussed in Section 1, using predefined question header templates allow the multiplication of generated questions. We use the same method, as shown in Table 1. An example of question header template is “Select the group items that belong to <Concept X>”, where <Concept X> is called the *missing concept*, and is instantiated during the question header generation by selecting a contract model concept from IFKB. The selected concept is called the *target concept*, and it will be used to retrieve/create the correct answer and the distractors in the next steps.

For example in Figure 4, the chosen question header template is “Select the group items that belong to <Concept X>”, and the target concept is “Salam Contract Model”.

TABLE I. EXCERPT OF QUESTION TEMPLATES

Types	Question Templates
DMCQ Type 1	<ul style="list-style-type: none"> Which of the following group items belong to <Concept X>? For the below choices, which one can be considered as part of <Concept X>? Select the group items that belong to <Concept X>
DMCQ Type 2	<ul style="list-style-type: none"> Fill in the blank for below diagram representing <Concept X>. For the diagram below representing <Concept X>, please fill in the blank. Fill in the blank with the correct answer for the diagram below representing <Concept X>.
DMCQ Type 3	<ul style="list-style-type: none"> Select the diagram that represents <Concept X>. Which of the following diagrams represent <Concept X>? Please select the diagram that represents <Concept X>.
DMCQ Type 4	<ul style="list-style-type: none"> Which of the below process is part of <Concept X> process? Select the diagram which is part of <Concept X> process Please select the diagram which is part of <Concept X> process.

For DMCQ Type 2, the question header also contains a modified diagram (see Figure 5), where parts of the original diagram have been randomly removed.

C. Correct Answer Generation

The correct answer of DMCQ Type 1 is generated by randomly choosing a subset of nodes/arcs from the original diagram. For DMCQ Types 2 and 3, the correct answer is straightforward, respectively the removed labels/nodes/arcs and the complete diagram. Finally, the correct answer of DMCQ Type 4 is a randomly chosen subset of the original diagram.

D. Distractors Generation

When generating the distractors, we have to consider two modifier strategies: labelling and structural strategies. They respectively concern the labels of the nodes and arcs,

and the structure of the contract model chosen as target concept. The labelling strategy is similar to existing class-based, property-based and terminology-based strategies [1], and consists in modifying the labels of nodes and/or arcs. The structural strategy modifies the correct contract model to create a distractor by adding new nodes and/or arcs, modifying direction of arcs, modifying the positioning of nodes, as well as removing nodes and/or arcs. By considering labelling as well as structural strategies, the number of possible generated distractors is growing immensely. The utilized strategies are summarized as follows:

- Strategy 1: Replacing a node label (resp. an arc label).
- Strategy 2: Inserting new nodes and/or arcs that do not belong to the correct contract model.
- Strategy 3: Modifying the direction and/or numbering of arcs.
- Strategy 4: Modifying the positioning of nodes by swapping nodes
- Strategy 5: deleting nodes and/or arcs belonging to the correct contract model.
- Strategy 6: Selecting a wrong contract model.

E. *Difficulty of Distractors for each Strategy*

In this research, we consider the generated distractors to be the core criterion for determining the question difficulty. The results of our preliminary investigation are as follows:

- In Strategy 1, the replacement label makes the question more difficult if its meaning is close to the replaced label, which is if it comes from a sibling or a direct parent in the hierarchy.
- In Strategy 2 (resp. 5), inserting (resp. deleting) n nodes/arcs makes the question easier as n grows bigger, since it totally alters the original diagram.
- Strategy 3 (resp. 4) introduces difficulty, as the diagram must be well understood in order to remember the direction as well as the numbering of the arcs (resp. the positioning of the nodes).
- Finally, Strategy 6 makes the question more difficult if the selected contract model is close in structure to the correct contract model.

IV. CONCLUSION

In this paper, we presented a methodology to automatically generate DMCQ from a knowledge base, and discussed our preliminary investigations on the difficulty of the generated questions. We introduced four types of DMCQs, and new generation strategies based on the structure of Islamic finance contract models in the form of diagrams.

In our future work, we plan to evaluate and extend our current prototype in several prospective directions: (1) evaluate the quality of the generated questions with regard to several criteria such as pedagogy; (2) pursue the investigation on question difficulty; (3) enrich the question templates; (3) extend the types of generated questions to free-hand drawing, and automatically assess the drawings.

REFERENCES

- [1] A. Papasalouros, K. Kanaris, and K. Kotis, "Automatic Generation Of Multiple Choice Questions From Domain Ontologies," presented at the IADIS e-Learning conference, Amsterdam, The Netherlands, 2008, pp.427-434.
- [2] M. Tasic and M. Cubric, "SeMCQ – Protégé Plugin for Automatic Ontology-Driven Multiple Choice Question Tests Generation", presented at the 11th International Protégé Conference, Amsterdam, The Netherlands, 2009.
- [3] M. Cubric and M. Tasic, "Towards automatic generation of e-assessment using semantic web technologies". In Proceedings of the 2010 International Computer Assisted Assessment Conference, University of Southampton, July, 2010.
- [4] A. Mamadolimova, N. Ambiah, and D. Lukose, "Modeling Islamic finance knowledge for contract compliance in Islamic banking", Proc. 15th international conference on Knowledge-based and intelligent information and engineering systems, Springer-Verlag, Volume Part III, 2011, pp. 346-355.
- [5] W. Al-Zuhayli, "Financial Transactions in Islamic Jurisprudence" (originally alFiqhu al-Islāmī wa adillatūhu). Translated into English by Mahmoud Amin El-Gamal. Syria: Dār al-Fikr, 2003.
- [6] Gruff: A Grapher-Based Triple-Store Browser for AllegroGraph. retrieved January, 2015, from: <http://franz.com/agraph/gruff/>.
- [7] Concepts and Abstract Syntax, retrieved January, 2015, from: <http://www.w3.org/TR/rdf11-concepts/>.
- [8] M. Al-Yahya, "Ontology-Based Multiple Choice Question Generation", The Scientific World Journal, 2014.
- [9] M. Teitsma, J. Sandberg, M. Maris, and B. Wielinga, "Using an Ontology to Automatically Generate Questions for the Determination of Situations", 22nd International Conference, DEXA, 2011.